Identifying Important Risk Factors for Survival in Patient With Systolic Heart Failure Using Random Survival Forests

Eileen Hsich, MD; Eiran Z. Gorodeski, MD, MPH; Eugene H. Blackstone, MD; Hemant Ishwaran, PhD; Michael S. Lauer, MD

Background—Heart failure survival models typically are constructed using Cox proportional hazards regression. Regression modeling suffers from a number of limitations, including bias introduced by commonly used variable selection methods. We illustrate the value of an intuitive, robust approach to variable selection, random survival forests (RSF), in a large clinical cohort. RSF are a potentially powerful extensions of classification and regression trees, with lower variance and bias.

Methods and Results—We studied 2231 adult patients with systolic heart failure who underwent cardiopulmonary stress testing. During a mean follow-up of 5 years, 742 patients died. Thirty-nine demographic, cardiac and noncardiac comorbidity, and stress testing variables were analyzed as potential predictors of all-cause mortality. An RSF of 2000 trees was constructed, with each tree constructed on a bootstrap sample from the original cohort. The most predictive variables were defined as those near the tree trunks (averaged over the forest). The RSF identified peak oxygen consumption, serum urea nitrogen, and treadmill exercise time as the 3 most important predictors of survival. The RSF predicted survival similarly to a conventional Cox proportional hazards model (out-of-bag C-index of 0.705 for RSF versus 0.698 for Cox proportional hazards model).

Conclusions—An RSF model in a cohort of patients with heart failure performed as well as a traditional Cox proportional hazard model and may serve as a more intuitive approach for clinicians to identify important risk factors for all-cause mortality. **(*Circ Cardiovasc Qual Outcomes*. 2011;4:39-45.)**

Key Words: heart failure ■ prognosis ■ statistics ■ survival analyses

Most heart failure survival models are based on multivariable Cox proportional hazard regression.[1–6] To prevent overfitting and achieve parsimony, analysts often identify statistically significant variables by methods such as stepwise regression or χ^2 statistical score ranking.[1,3,7,8] These methods yield variable results, and have been criticized for creating bias.[9] In addition, from the point of view of clinicians, regression modeling and variable selection appear to occur within a computer's "black box."

Statistical methods like classification and regression trees may be intuitive for clinicians, because they illustrate the importance and relationship of variables with a single young tree that has few branches.[10] However, classification and regression trees suffer from high variance and poor performance,[11–13] which leads to instability. Random survival forests (RSF) modeling is a new statistical method that grows numerous mature trees with many branches.[14] RSF reduce variance and bias by using all variables collected and by automatically assessing for nonlinear effects and complex interactions. They are a direct extension of the random forest, which has been successfully used in clinical studies[15–18] and, in some cases, shown to outperform classical statistical methods.[18,19]

We used RSF to illustrate an intuitive and powerful approach for identifying important risk factors for survival in 2231 patients with systolic heart failure who underwent cardiopulmonary stress testing at the Cleveland Clinic. Variables with relatively high importance are near the tree trunks.[20] We also compared the results of RSF to our previously published Cox proportional hazard model for predictive accuracy of the model and for selection of important risk factors for all-cause mortality.[21]

Methods

Data Source

The design of this observational prospective study has been previously published.[21] The cohort consisted of all adult patients at the Cleveland Clinic with left ventricular ejection fraction <40% who underwent cardiopulmonary stress testing between August 1997 and April 2007 using a modified Naughton protocol, the most common

Received January 20, 2010; accepted September 20, 2010.

From the Heart and Vascular Institute (E.H., E.Z.G., E.H.B.), Department of Quantitative Health Sciences (E.H.B., H.I.), and Case Western Reserve University School of Medicine (E.H., E.H.B.), Cleveland, Ohio; and Division of Cardiovascular Sciences (M.S.L.), National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Md.

Correspondence to Michael S. Lauer, MD, Director, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Rockledge Center II, 6701 Rockledge Dr, Room 8128, Bethesda, MD 20892. E-mail lauerm@nhlbi.nih.gov

Circ Cardiovasc Qual Outcomes **is available at http://circoutcomes.ahajournals.org** **DOI: 10.1161/CIRCOUTCOMES.110.939371**

protocol used in our laboratory for heart transplantation evaluation. Patients were excluded if they were aged <18 years or had no U.S. Social Security number. Left ventricular ejection fraction was assessed by echocardiogram, left ventriculography, or ECG-gated SPECT imaging. If >1 stress test was performed on an individual, only the first was used in this analysis. Demographic information, height and weight directly measured, medications, and stress test results were entered into our electronic database at the time of stress testing.

WHAT IS KNOWN

- Classic regression models have serious limitations, including "black box" methods for determining which variables most strongly predict outcome.
- The technique of random survival forests (RSF) is a robust, computer-based algorithm that yields unbiased assessments of variable importance.
- RSF and related techniques have been primarily used in fields outside of clinical medicine.

WHAT THIS STUDY ADDS

- We have shown that RSF can be used to select the most important variables predictive of mortality in patients with severe heart failure.

The results of exercise stress testing were recorded on a MedGraphic cardiopulmonary system (St Paul, Minn). Heart rate, blood pressure, respiratory rate, oxygen consumption ($\dot{V}O_2$), carbon dioxide production, minute ventilation, and tidal volume were obtained every 30 seconds at rest, during exercise, and during recovery. Exercise stress testing was symptom limited, and total duration of exercise was measured to the nearest second. Serum laboratory tests within 3 months were included, and only the tests closest in time to the stress test were considered. As we discussed previously,[21] laboratory tests before October 1999 were systematically missing from our electronic database; therefore, we used informed imputation to fill in 10% of serum glucose, serum urea nitrogen (BUN), creatinine, and sodium values and 15% of hemoglobin values. No other data were missing either systematically or at random, precluding any need for multiple imputation.[22] Glomerular filtration rate was estimated using the Cockcroft-Gault equation.[23] The study was approved by the Institutional Review Board at the Cleveland Clinic, and informed consent was waived because all data were collected and recorded as part of routine clinical care.

Study Variables

The following variables were assessed for prognostic value: sex, age, body mass index (kg/m^2), current tobacco use, insulin-treated diabetes, noninsulin-treated diabetes, coronary artery disease, previous myocardial infarction, previous coronary artery bypass graft surgery, previous percutaneous coronary intervention, implantable cardioverter-defibrillator (ICD), pacemaker, β-blocker, angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, potassium-sparing diuretics, antiarrythmics, anticoagulation, aspirin, digoxin, nitrates, vasodilators, loop diuretics, thiazide diuretics, statins, nondihydropyridine calcium channel blocker, dihydropyridine calcium channel blocker, resting heart rate (beats per minute), resting systolic blood pressure (mm Hg), left ventricular ejection fraction, peak $\dot{V}O_2$ (mL/kg per min) peak respiratory exchange ratio, treadmill exercise time, serum sodium (mmol/L), creatinine clearance (mL/min), BUN (mg/dL), serum hemoglobin (g/dL), and serum glucose (mg/dL).

End Points

The primary end point was all-cause death. Mortality data were obtained by linking our database with the U.S. Social Security Administration Death Index, which we previously reported to have a sensitivity of 97%.[24]

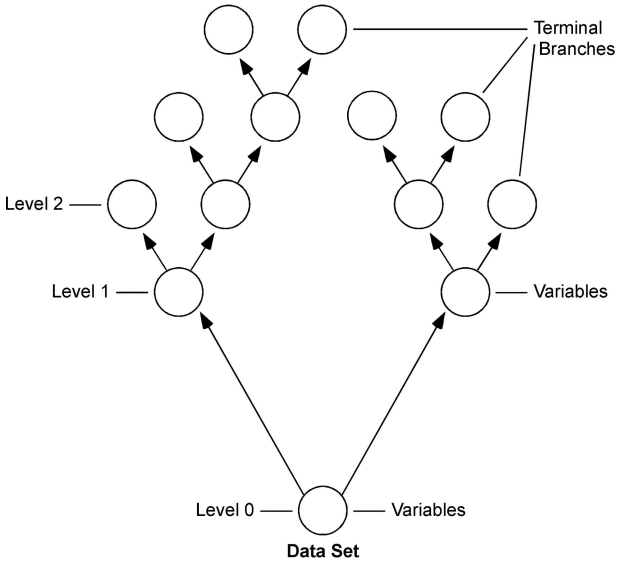

Figure 1. Example of a random tree. A bootstrap sample of patients from the original data set is used to create a random tree. At the tree trunk (or root node), a random set of variables is chosen to be candidates, and the most predictive variable for survival among those is identified. Node levels are numbered based on their relative distance to the trunk of the tree (ie, 0, 1, 2). Splitting of nodes to create the tree continues until terminal nodes have few distinct deaths.

Statistical Analysis

Sex-specific baseline characteristics were reported, with continuous variables expressed as means±SD and categorical variables as frequencies. Random survival analysis used all-cause mortality for the outcome.[24] Thirty-nine variables in 2231 patients were used for the analysis. A survival forest of 2000 survival trees was constructed.

Figure 1 demonstrates how we build a single random tree. We start by choosing a bootstrap sample of patients from the original cohort. At each branch, a random set of variables are chosen as candidates to split the branch into 2 other branches, and the variable maximizing the log-rank statistic[25] using 3 randomly selected split points was used for splitting. The number of variables assessed at each branch was the square root of the total number of variables. Branch levels are numbered on the basis of their relative distance from the tree trunk (ie, 0, 1, 2). Splitting of branches to create the tree continues as long as possible until terminal branches have no fewer than 3 deaths.

An RSF is generated by creating 2000 trees. The most important variables are identified as those that most frequently split the branches near the tree trunks. There are no prespecified assumptions regarding variables, and randomization is introduced into this model by both random bootstrap sampling of patients from the original cohort and random sampling of variables for each tree branch. Importance of a variable is assessed by minimal depth from the tree trunk.[14] To illustrate this concept, we show in Figure 2 a random tree with color coding of maximal subtrees. A maximal subtree for a variable v is the largest subtree whose lowest branch is split using v. The shortest distance from the tree trunk to the branch level of the closest maximal subtree of v is the minimal depth of v. For example, in Figure 2, exercise time splits the tree trunk and has a minimal depth of 0, whereas BUN is the 2 green subtrees with a minimal depth of 2. The most predictive variables for the cohort are defined as those whose minimal depth (averaged over the forest) is smaller than the mean minimal depth determined under the null hypothesis of no effect.[20]

Figure 2. Illustration of minimal depth of a variable in a random tree from our 2000-tree forest. Highlighted are the 3 top variables: peak $\dot{V}O_2$ (red), BUN (green), and exercise time (yellow). Depth of a node is indicated by numbers 0, 1, 2, and 3 to 8. The minimal depths are 0, 1, and 2 for exercise time, peak $\dot{V}O_2$, and BUN, respectively. AICD indicates automated implantable cardioverter-defibrillator; CABG, coronary artery bypass graft; LVEF, left ventricular ejection fraction; PCI, percutaneous coronary intervention.

Prediction accuracy for RSF was assessed by Harrell C-index using out-of-bag (OOB) data. The OOB method involves obtaining bootstrap samples from the original cohort and using each sample to compute a prediction model. Each bootstrap sample left out about one-third of the data, which was referred to as the OOB data. The C-index was calculated using an OOB ensemble constructed with the 2000 OOB data sets produced by the 2000 bootstrap samples used in deriving the forest.

A nonparsimonious Cox proportional hazards model was constructed as previously described[21] and compared with the RSF model for predictive accuracy of the model and for selection of important risk factors for all-cause mortality. Briefly, the proportional hazards assumption was tested by scaled Schoenfeld residuals and inspection of hazard ratio plots. Possible nonlinear associations for the Cox proportional hazards model were tested with restricted cubic splines, and possible interactions were tested. Prediction accuracy for the Cox proportional hazards model was assessed by Harrell C-index OOB data.[21]

All analyses were performed with SAS version 9.1.3 (SAS Institute; Cary, NC) and R version 2.6.2 (www.R-project.org). RSF were implemented using the "RandomSurvivalForest" R-package, freely available through the Comprehensive R Archive Network distribution system (http://cran.r-project.org/web/packages/randomSurvivalForest/index.html).

Results

Our cohort consisted of 2231 patients, including 602 (27%) women and 1629 (73%) men. There were 155 women (26% of female cohort) and 587 (36% of male cohort) men who died during a mean follow-up of 5 years (maximum for survivors, 11 years).

The Table shows the baseline characteristics of the cohort according to sex. Our patients had advanced disease with low systolic blood pressure, low peak $\dot{V}O_2$, and low left ventricular ejection fraction. Most patients received angiotensin-converting enzyme inhibitors or angiotensin receptor blockers, and >60% received β-blockers.

Figure 3 shows 6 randomly chosen trees from the 2000-tree forest. The 3 most important variables among these trees are color coded blue for treadmill exercise time, red for peak $\dot{V}O_2$, and green for BUN. These colors appear on almost every tree and are found near the tree trunks, demonstrating their relative importance.

Figure 4 shows all 39 variables and plots their minimal depth. The horizontal line separates the 10 predictive variables from the remaining nonpredictive variables. The 3

Table. Sex-Specific Baseline Characteristics

Variable	All (N=2231)	Women (n=602)	Men (n=1629)
Age, y	54±11	52±11	55±11
Body mass index, kg/m^2	28±6	28±6	29±5
Current smokers	459 (21)	117 (19)	342 (21)
Diabetes, insulin treated	215 (10)	53 (9)	162 (10)
Diabetes, noninsulin treated	350 (16)	92 (15)	258 (16)
Coronary artery disease	906 (41)	127 (21)	779 (48)
Previous MI	279 (13)	43 (7)	236 (14)
Previous CABG	594 (27)	64 (11)	530 (33)
Previous PCI	476 (21)	75 (12)	401 (25)
ICD	647 (29)	147 (24)	500 (31)
Pacemaker	502 (23)	113 (19)	389 (24)
Medication use			
β-blocker	1429 (64)	387 (64)	1042 (64)
ACE inhibitor	1711 (77)	431 (72)	1280 (79)
Angiotensin receptor blocker	290 (13)	99 (16)	191 (12)
Potassium-sparing diuretics	649 (29)	203 (34)	446 (27)
Antiarrythmic	509 (23)	90 (15)	419 (26)
Anticoagulation	899 (40)	210 (35)	689 (42)
Aspirin	1038 (47)	230 (38)	808 (50)
Digoxin	1570 (70)	424 (70)	1146 (70)
Nitrates	739 (33)	153 (25)	586 (36)
Vasodilators	136 (6)	27 (4)	109 (7)
Loop diuretics	1880 (84)	498 (83)	1382 (85)
Thiazide diuretics	279 (13)	77 (13)	202 (12)
Statin	850 (38)	172 (29)	678 (42)
Calcium channel blocker, nondihydropyridine	16 (1)	4 (1)	12 (1)
Calcium channel blocker, dihydropyridine	99 (4)	15 (2)	84 (5)
Resting heart rate, beats/min	76±14	78±14	76±14
Resting systolic blood pressure, mm Hg	111±18	110±18	111±18
LVEF, (%)	20±7	21±7	20±7
Peak $\dot{V}o_2$, mL/kg per min	16±5	16±4	17±5
Peak respiratory exchange ratio	1.08±0.12	1.05±0.13	1.09±0.11
Treadmill exercise time, s	503±221	476±204	513±226
Serum sodium, mmol/L	139±3	140±3	139±3
Creatinine clearance, mg/min	91±43	85±44	93±43
BUN, mg/dL	25±13	23±12	26±13
Serum hemoglobin, g/dL	14±1	13±1	14±1
Serum glucose, mg/dL	109±43	105±40	111±43

Data are presented as no. (%) or mean±SD. Treadmill exercise time=maximal interval for phase 2 (seconds)±SD (seconds). ACE indicates angiotensin-converting enzyme; CABG, coronary artery bypass graft; LVEF, left ventricular ejection fraction; MI, myocardial infarction; PCI, percutaneous coronary intervention.

variables on the extreme left are peak $\dot{V}O_2$, BUN, and treadmill exercise time and are easily seen to be the most-predictive variables. These variables are similar to what was found in our previously published Cox proportional hazard model analysis but in a different relative order (ie, peak $\dot{V}O_2$, treadmill exercise time, and BUN).[21]

Figure 5 displays how the RSF model shows interaction among these 3 most important variables and 5-year predicted survival. Patients with the highest peak $\dot{V}O_2$ and longest treadmill exercise time have the best survival (first row, last column), and most had low BUN. Survival was worst for patients with the lowest peak $\dot{V}O_2$ and shortest treadmill time (last row, first column) and further depended on small changes in BUN between 20 and 40 mg/dL. In this group, 5-year predicted survival was about 70% for those with a BUN of 20 mg/dL but only about 50% for those with a BUN of 40 mg/dL. Survival did not change much for those with BUN >40 mg/dL. Among those with the lowest peak $\dot{V}O_2$ (first column) survival depended more on BUN than on treadmill time. For those with the shortest exercise time (last row), survival also was very dependent on BUN. It is important to note that these interactions and nonlinear relationships were identified by the forest and not prespecified by the analyst.

Figure 6 is similar to Figure 5 but provides the added dimension of β-blockers. Five-year predicted survival was worse for all groups not taking β-blockers at the time of the cardiopulmonary stress testing. The greatest differences in survival were among patients with a BUN >40 mg/dL.

We compared the RSF model to a Cox proportional hazard model. Model discrimination was similar using RSF analysis with an OOB C-index of 0.705 compared to our previously published nonparsimonious Cox proportional hazard model with a C-index of 0.698.[21] Using the 10 most important variables selected by the RSF model to create another Cox proportional hazard model, the C-index for this simplified Cox proportional hazard model was comparable to the nonparsimonious Cox proportional hazard model that included >30 variables (C-index, 0.699 versus 0.698).

Discussion

RSF identified peak $\dot{V}O_2$, BUN, and treadmill exercise time as the top-3 most important predictors of survival in our cohort of 2231 ambulatory patients with systolic heart failure who underwent cardiopulmonary stress testing at the Cleveland Clinic. These variables are similar to what was found in our previously published Cox proportional hazard model analysis but in a different relative order.[21] The method used to determine the most important predictors for RSF is easy for clinicians to understand and visualize because important predictor variables are located at the tree trunks of the forest, which can be color coded for easy identification. In addition, RSF predicted survival as well as the conventional Cox proportional hazard model did (OOB C-index for RSF was 0.705 compared with a C-index for a nonparsimonious Cox proportional hazard model of 0.698). Variable selection by RSF also was used to create a simplified Cox proportional hazard model that performed like a nonparsimonious Cox proportional hazard model constructed with >3 times the number of variables.[21]

There are 4 advantages to using RSF. First, the RSF method is intuitive because important variables to predict survival can be identified by inspecting the tree trunks and

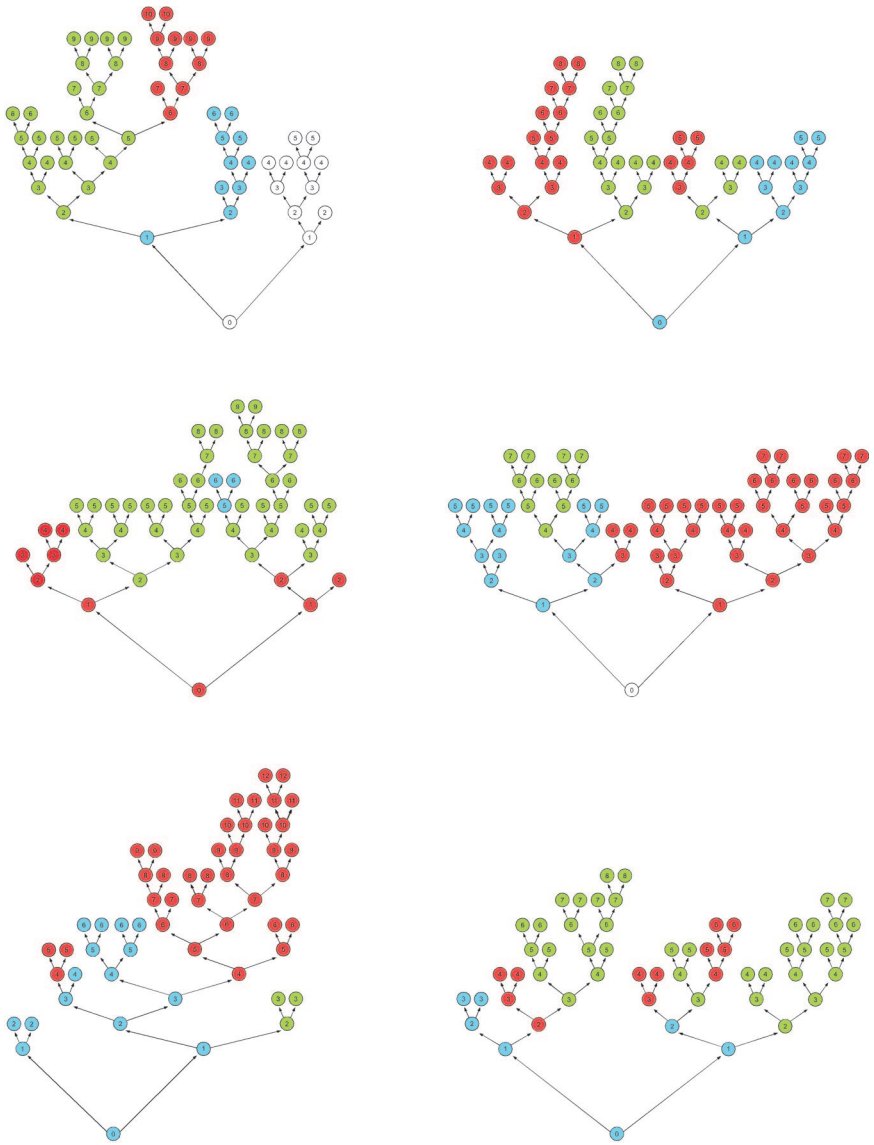

Figure 3. Illustration of 6 random trees from our 2000-tree forest. The 3 most important variables among these trees are color coded blue for treadmill exercise time, red for peak $\dot{V}O_2$, and green for BUN.

simplified in a figure plotting the minimal depth of a variable from the tree trunk. Second, RSF do not require analysts to know in advance the relationship (ie, linear, nonlinear) of a variable over time or to choose the best equation to transform nonlinear covariates. Third, the complex interactions among multiple variables can be easily understood with RSF, using plots such as those shown in Figures 5 and 6. Finally, the overall accuracy of an RSF model is at least comparable to standard methodologies.[14]

RSF is a new, robust extension of random forest, a well-known and highly used machine learning method, and has been used successfully in several applied settings, including staging esophageal cancer[26,27] and genomics.[28] Machine learning involves use of computers to generate "automatic techniques for learning to make accurate predictions based on past observations."[29] All variables collected can be used for the survival analysis, and the method for variable selection is intuitive and has been shown to outperform parametric methods as well as other state-of-the-art machine learning methodologies.[20] RSF do not rely on *P* values, and analysts do not need to select important variables in advance with methods like stepwise regression, inspect for residuals, or include interactions. Several large studies (using simulations and real data) have now compared RSF to other methods, including Cox regression, and these have shown RSF to be consistently better than, or at least as good as, competing methods.[14,18] Since the introduction of random forest to the machine learning community almost 10 years ago,[30] there have been efforts to document its empirical performance. Our results confirm what has generally been found: random forest produces accurate prediction.[14,18] Our study, using a large cohort of consecutive patients with heart failure with very low loss of follow-up, showed that the RSF model was at least as good as Cox regression with respect to survival prediction. More studies are needed to compare RSF to Cox regression to further document their performance in clinical settings.

The major limitation of our study is that we have not validated either RSF or our Cox proportional hazard model with an external cohort from another advanced heart failure center. Although RSF effectively validate the model by

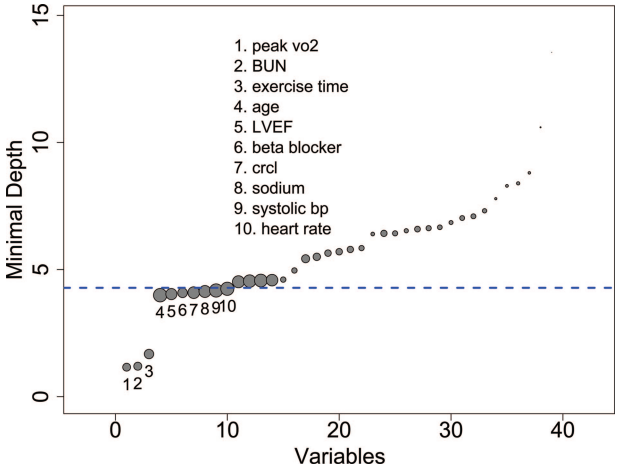

Figure 4. Minimal depth (variable importance) from RSF analysis. Horizontal line is threshold for filtering variables. All variables below the line are predictive. The diameter of each circle in the plot is proportional to the forest-averaged number of maximal subtrees for that variable. bp indicates blood pressure; crcl, creatinine clearance; LVEF, left ventricular ejection fraction.

creating trees with a random group of patients and variables, the model is still deriving these trees from the original data set, and performance with an external cohort will need to be assessed. Other limitations include the fact that more variables could be included and that variables commonly accepted as predictors of survival, such as serum B-type natriuretic peptides, were not routinely obtained at our center between 1997 and 2007. Biventricular pacemakers also were not reported separately during database entry, but most were identified in the ICD category because at our institution, biventricular pacemakers were almost always implanted with an ICD. We cannot account for variables that change with time that may affect mortality, and we plan further work on

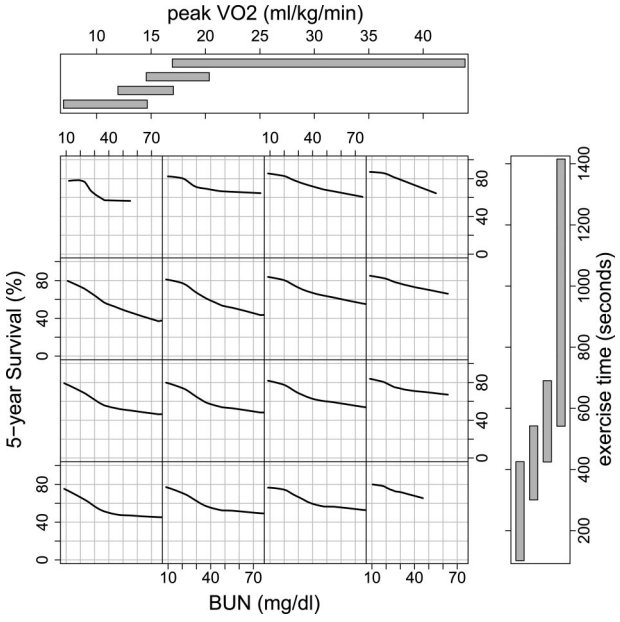

Figure 5. RSF-estimated 5-year survival as a function of BUN, exercise time, and peak $\dot{V}O_2$. Smoothed curves are loess curves of the estimated survival for each individual.

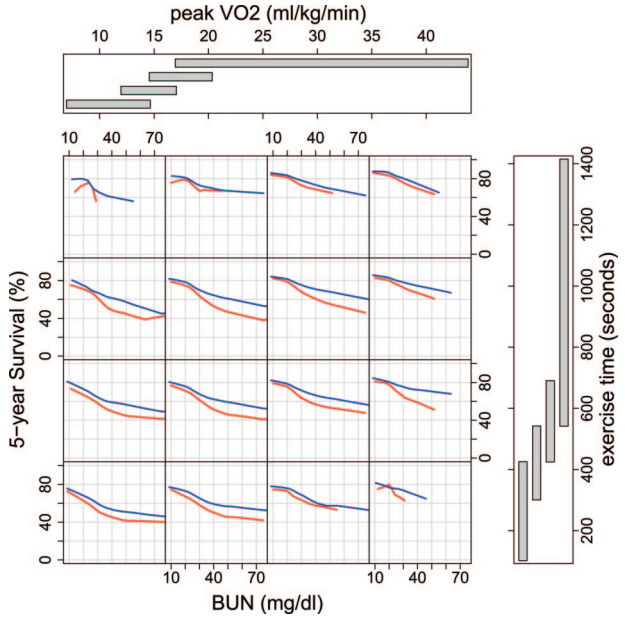

Figure 6. RSF-estimated 5-year survival as a function of BUN, exercise time, and peak $\dot{V}O_2$ for patients taking and not taking β-blockers at the time of first cardiopulmonary stress test at the Cleveland Clinic. Blue curves identify patients taking β-blockers, and red curves identify those not taking β-blockers. Smoothed curves are loess curves of the estimated survival for each individual.

developing capabilities to analyze time-dependent covariates. However, the majority of the limitations described herein, with the exception of the need to externally validate, are what limit our survival model from possibly being better than other survival models, but they do not prevent a fair comparison of RSF to a Cox proportional hazard model.

In summary, we found in a large, single-center cohort of patients with severe systolic heart failure that RSF identified similar risk factors to predictors of all-cause mortality and that an RSF model performed as well as the traditional Cox proportional hazard model. The RSF method holds promise as an intuitive approach for variable selection and as a way to eliminate the mistrust in the black box approach to statistical analysis.

Sources of Funding

This work was supported in part by the Health Resources and Services Administration contract 234-2005-370011C; American Heart Association (AHA) Scientist Development Grant 0730307N; and the National Heart, Lung, and Blood Institute (NHLBI) CAN #8324207 and contract HHSN268200800026C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the AHA, NHLBI, or Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

Disclosures

None.

References

1. Aaronson KD, Schwartz JS, Chen TM, Wong KL, Goin JE, Mancini DM. Development and prospective validation of a clinical index to predict

survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation*. 1997;95:2660–2667.
2. Brophy JM, Dagenais GR, McSherry F, Williford W, Yusuf S. A multivariate model for predicting mortality in patients with heart failure and systolic dysfunction. *Am J Med*. 2004;116:300–304.
3. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*. 2006;113:1424–1433.
4. Mullens W, Abrahams Z, Skouri HN, Taylor DO, Starling RC, Francis GS, Young JB, Tang WH. Prognostic evaluation of ambulatory patients with advanced heart failure. *Am J Cardiol*. 2008;101:1297–1302.
5. Stempfle HU, Alt A, Stief J, Siebert U. The Munich score: a clinical index to predict survival in ambulatory patients with chronic heart failure in the era of new medical therapies. *J Heart Lung Transplant*. 2008;27:222–228.
6. Zugck C, Kruger C, Kell R, Korber S, Schellberg D, Kubler W, Haass M. Risk stratification in middle-aged patients with congestive heart failure: prospective comparison of the Heart Failure Survival Score (HFSS) and a simplified two-variable model. *Eur J Heart Fail*. 2001;3:577–585.
7. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006;355:2631–2639.
8. Rautaharju PM, Kooperberg C, Larson JC, LaCroix A. Electrocardiographic abnormalities that predict coronary heart disease events and mortality in postmenopausal women: the Women's Health Initiative. *Circulation*. 2006;113:473–480.
9. Snedecor GW, Cochran WG. *Statistical methods*. 8th ed. Ames, IA: Iowa State University Press; 1989.
10. Breiman L, Friedman JH, Olsthen RA, Stone J. *Classification and Regression Trees*. Monterey, CA: Wadsworth International; 1984.
11. Austin PC, Tu JV, Lee DS. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J Clin Epidemiol*. 2010;63:1145–1155.
12. Breiman L. Bagging predictors. *Machine Learning*. 1996;24:123–140.
13. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat*. 1996;24:2350–2383.
14. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841–860.
15. Ward MM, Pajevic S, Dreyfuss J, Malley JD. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests. *Arthritis Rheum*. 2006;55:74–80.
16. Heidema AG, Feskens EJ, Doevendans PA, Ruven HJ, van Houwelingen HC, Mariman EC, Boer JM. Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. *Genet Epidemiol*. 2007;31:910–921.
17. Mamyrova G, O'Hanlon TP, Monroe JB, Carrick DM, Malley JD, Adams S, Reed AM, Shamim EA, James-Newton L, Miller FW, Rider LG. Immunogenetic risk and protective factors for juvenile dermatomyositis in Caucasians. *Arthritis Rheum*. 2006;54:3979–3987.
18. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet*. 2004;5:32.
19. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*. 2006;63:490–500.
20. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High dimensional variable selection for survival data. *J Am Stat Assoc*. 2010;105:205–217.
21. Hsich E, Gorodeski EZ, Starling RC, Blackstone EH, Ishwaran H, Lauer MS. Importance of treadmill exercise time as an initial prognostic screening tool in patients with systolic left ventricular dysfunction. *Circulation*. 2009;119:3189–3197.
22. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer; 2001:49.
23. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. *Nephron*. 1976;16:31–41.
24. Nishime EO, Cole CR, Blackstone EH, Pashkow FJ, Lauer MS. Heart rate recovery and treadmill exercise score as predictors of mortality in patients referred for exercise ECG. *JAMA*. 2000;284:1392–1398.
25. Segal MR. Regression trees for censored-data. *Biometrics*. 1988;44:35–47.
26. Ishwaran H, Blackstone EH, Apperson-Hansen C, Rice TW. A novel approach to cancer staging: application to esophageal cancer. *Biostatistics*. 2009;10:603–620.
27. Rizk NP, Ishwaran H, Rice TW, Chen LQ, Schipper PH, Kesler KA, Law S, Lerut TEMR, Reed CE, Salo JA, Scott WJ, Hofstetter WL, Watson TJ, Allen MS, Rusch VW, Blackstone EH. Optimum lymphadenectomy for esophageal cancer. *Ann Surg*. 2010;251:46–50.
28. Weichselbaum RR, Ishwaran H, Yoon T, Nuyten DS, Baker SW, Khodarev N, Su AW, Shaikh AY, Roach P, Kreike B, Roizman B, Bergh J, Pawitan Y, van de Vijver MJ, Minn AJ. An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. *Proc Natl Acad Sci U S A*. 2008;105:18490–18495.
29. Schapire RE. The boosting approach to machine learning: an overview. Available at: http://74.125.155.132/scholar?q=cache:YirIIUAAd_kJ:scholar.google.com/+definition+machine+learning+method&hl=en&as_sdt=20000000&as_vis=1. Accessed August 24, 2010.
30. Breiman L. Random forests. *Machine Learning*. 2001;45:5–32.