Analysis of Machine Learning Techniques for Heart Failure Readmissions

Bobak J. Mortazavi, PhD; Nicholas S. Downing, MD; Emily M. Bucholz, MD, PhD;
Kumar Dharmarajan, MD, MBA; Ajay Manhapra, MD; Shu-Xia Li, PhD;
Sahand N. Negahban, PhD*; Harlan M. Krumholz, MD, SM*

Background—The current ability to predict readmissions in patients with heart failure is modest at best. It is unclear whether machine learning techniques that address higher dimensional, nonlinear relationships among variables would enhance prediction. We sought to compare the effectiveness of several machine learning algorithms for predicting readmissions.

Methods and Results—Using data from the Telemonitoring to Improve Heart Failure Outcomes trial, we compared the effectiveness of random forests, boosting, random forests combined hierarchically with support vector machines or logistic regression (LR), and Poisson regression against traditional LR to predict 30- and 180-day all-cause readmissions and readmissions because of heart failure. We randomly selected 50% of patients for a derivation set, and a validation set comprised the remaining patients, validated using 100 bootstrapped iterations. We compared C statistics for discrimination and distributions of observed outcomes in risk deciles for predictive range. In 30-day all-cause readmission prediction, the best performing machine learning model, random forests, provided a 17.8% improvement over LR (mean C statistics, 0.628 and 0.533, respectively). For readmissions because of heart failure, boosting improved the C statistic by 24.9% over LR (mean C statistic 0.678 and 0.543, respectively). For 30-day all-cause readmission, the observed readmission rates in the lowest and highest deciles of predicted risk with random forests (7.8% and 26.2%, respectively) showed a much wider separation than LR (14.2% and 16.4%, respectively).

Conclusions—Machine learning methods improved the prediction of readmission after hospitalization for heart failure compared with LR and provided the greatest predictive range in observed readmission rates. **(*Circ Cardiovasc Qual Outcomes*. 2016;9:629-640. DOI: 10.1161/CIRCOUTCOMES.116.003039.)**

Key Words: computers ■ heart failure ■ machine learning ■ meta-analysis ■ patient readmission

High rates of readmission after hospitalization for heart failure impose tremendous burden on patients and the healthcare system.[1–3] In this context, predictive models facilitate identification of patients at high risk for hospital readmissions and potentially enable direct specific interventions toward those who might benefit most by identifying key risk factors. However, current predictive models using administrative and clinical data discriminate poorly on readmissions.[4–8] The inclusion of a richer set of predictor variables encompassing patients' clinical, social, and demographic domains, while improving discrimination in some internally validated studies,[9] does not necessarily markedly improve discrimination,[10] particularly in the data set to be considered in this work. This richer set of predictors might not contain the predictive domain of variables required, but does represent a large set of data not routinely collected in other studies.

Another possibility for improving models, rather than simply adding a richer set of predictors, is that prediction might improve with methods that better address the higher order interactions between the factors of risk. Many patients may have risk that can only be predicted by modeling complex relationships between independent variables. For example, no available variable may be adequately explanatory; however,

Received May 26, 2016; accepted October 17, 2016.

From the Section of Cardiovascular Medicine, Department of Internal Medicine (B.J.M., N.S.D., E.M.B., K.D., H.M.K.), Department of Psychiatry and the Section of General Medicine, Department of Internal Medicine (A.M.), and Robert Wood Johnson Foundation Clinical Scholars Program, Department of Internal Medicine, Yale School of Medicine, and Department of Health Policy and Management (H.M.K.), Yale School of Public Health, New Haven, CT; Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT (B.J.M., N.S.D., E.M.B., K.D., S.-X.L., H.M.K.); and Department of Statistics, Yale University, New Haven, CT (B.J.M., S.N.N.).

Current address for N.S.D.: Brigham and Women's Hospital, Boston, MA.

Current address for E.M.B.: Boston Children's Hospital, Boston, MA.

*Drs Negahban and Krumholz contributed equally as senior authors to this work.

This article was handled independently by Javed Butler, MD, MPH, as a Guest Editor. The editors had no role in the evaluation or in the decision about its acceptance.

The Data Supplement is available at http://circoutcomes.ahajournals.org/lookup/suppl/doi:10.1161/CIRCOUTCOMES.116.003039/-/DC1.

Correspondence to Harlan M. Krumholz, MD, SM, Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, 1 Church St, Suite 200, New Haven, CT 06510. E-mail harlan.krumholz@yale.edu

WHAT IS KNOWN

- Prediction models for readmissions in heart failure are often modest, at best, at discriminating between readmitted and nonreadmitted patients.

WHAT THE STUDY ADDS

- This study compares popular machine learning methods for comparison against traditional techniques.
- This study introduces models that can select variables for us from a larger, comprehensive data set.
- This study shows the improvements in using machine learning methods and a comprehensively collected data set, to help readers understand factors that contribute to readmission and limitations to current methods in readmission prediction

interactions between variables may provide the most useful information for prediction.

Modern machine learning (ML) approaches can account for nonlinear and higher dimensional relationships between a multitude of variables that could potentially lead to an improved explanatory model.[11,12] Many methods have emerged from the ML community that can construct predictive models using many variables and their rich nonlinear interactions.[13–15] Three widely used ML approaches may provide utility for readmission prediction: random forest (RF), boosting, and support vector machines (SVM). The primary advantage of ML methods is that they handle nonlinear interactions between the available data with similar computational load. RF involves the creation of multiple decision trees[16] that sort and identify important variables for prediction.[14,17] Boosting algorithms harness the power of weaker predictors by creating combined and weighted predictive variables.[18,19] This technique has been applied to preliminary learning work with electronic health records.[20] SVM is a methodology that creates clearer separation of classes of variables using nonlinear decision boundaries, or hyperplanes, from complex, multidimensional data in possibly infinite dimensional spaces.[21–26]

In this study, we tested whether these ML approaches could predict readmissions for heart failure more effectively than traditional approaches using logistic regression (LR).[7,9,27,28] We tested these strategies using data that included detailed clinical and sociodemographic information collected during the Tele-HF trial (Telemonitoring to Improve Heart Failure Outcomes),[29] a National Institutes of Health–sponsored randomized clinical trial to examine the effect of automated telemonitoring on readmission after hospitalization for heart failure.[10,29,30] We further tested whether the ML techniques could be improved when used hierarchically where outputs of RF were used as training inputs to SVM or LR. We evaluated these approaches by their effect on model discrimination and predictive range to understand ML techniques and evaluate their effectiveness when applied to readmissions for heart failure.

Methods

Data Source

Data for this study were drawn from Tele-HF, which enrolled 1653 patients within 30 days of their discharge after an index hospitalization for heart failure.[30] In addition to the clinical data from the index admission, Tele-HF used validated instruments to collect data on patients' socioeconomic, psychosocial, and health status. This study collected a wide array of instrumented data, from comprehensive, qualitative surveys to detailed hospital examinations, including many pieces of data not routinely collected in practice, providing a larger exploratory set of variables that might provide information gain.

The primary outcome was all-cause readmission or mortality within 180 days of enrollment.[29] A committee of physicians adjudicated each potential readmission to ensure that the event qualified as a readmission rather than another clinical encounter (eg, emergency department visit) and to determine the primary cause of the readmission. The comprehensive nature in which outcomes were tracked and determined across the various sites makes this a well-curated data set that can potentially leverage this information where other trials may not, as readmissions often occur at other institutions external to the study network.[31] Results of the Tele-HF study revealed that outcomes were not significantly different between the telemonitoring and control arms in the primary analysis.[29]

Analytic Sample Selection

For this study, we included all patients whose baseline interviews were completed within 30 days of hospital discharge to ensure that the information was reflective of the time of admission. Of the 1653 enrolled patients, we excluded 36 who were readmitted or died before the interview, 574 whose interviews were completed after 30 days from discharge, and 39 who were missing data on >15 of the 236 baseline features to create a study sample of 1004 patients for the 30-day readmission analysis set. To create the 180-day readmission analysis set, we further excluded 27 patients who died before the end of the study and had no readmission events, leaving 977 patients in the sample.

Feature Selection

We used 472 variables (called features in the ML community) for input. We first gathered the full 236 baseline patient characteristics available in Tele-HF.[10,30] This set included data extracted from medical record abstractions, hospital laboratory results, physical examination information, as well as quality of life, socioeconomic, and demographic information from initial patient surveys. The polarity of qualitative and categorical questions was altered if necessary to ensure that the lowest values reflect a strongly negative answer or missing data, and the highest values correspond to strongly positive answers (Data Supplement). In addition, we created dummy variables for each of the 236 features to indicate whether the value was missing or not (0 and 1 values).

Definition of Outcomes

We developed methods to predict 4 separate outcomes: (1) 30-day all-cause readmission; (2) 180-day all-cause readmission; (3) 30-day readmission because of heart failure; and (4) 180-day readmission because of heart failure. We trained predictive models for each of these outcomes and compared them with each other.

Predictive Techniques

We built models using both traditional statistical methods and ML methods to predict readmission, and compared model discrimination and predictive range of the various techniques. For traditional statistical methods, we used an LR model, and a Poisson regression. Three ML methods—RF, boosting, and SVM—were used for readmission prediction.

Logistic Regression

A recent study from Tele-HF used random survival forest methods to select from a comprehensive set of variables for Cox regression analysis.[10] The article had a comprehensive evaluation of the most predictive variables in the Tele-HF data set, using various techniques and various validation strategies. Using the variables selected in the article would provide the most accurate representation of an LR model on the Tele-HF data set for comparison purposes. Therefore, we used the same selected variables for our current study to compare model performance, as the current analysis is concerned with finding improved analytic algorithms for predicting 30-day readmissions rather than primarily with variable selection.

To verify that this LR model, which leverages the comprehensive variable selection technique in previous work,[10] was the best model for comparison, we compared its performance against other LR models built using varied feature selection techniques. The first model selected variables for LR by lasso regularization, the next a forward, stepwise feature selection technique based on each variable's likelihood ratio. All models were validated using 100 bootstrapped iterations; the former could not find a set of predictive variables, the latter varied in features chosen but selected, on average, only 5 variables. The model built using features selected from the work of Krumholz et al[10] outperformed the other techniques, when comparing mean C statistics. For a further detailed discussion of the other LR models built, we refer readers to the Data Supplement.

Comparison Techniques

Given the flexibility of nonlinear methods, the complexity of the desired models might overwhelm the available data, resulting in overfitting. Although all the available variables can be used in ML techniques such as RF and boosting, which are robust to this overfitting, we may require some form of feature selection to help prevent overfitting in less robust techniques like SVM.[22] Further explanation of the predictive techniques, as well as the evaluation method, is provided in the Data Supplement and additionally in the textbook by Hastie et al.[16]

Hierarchical Methods

To overcome the potential for overfitting in LR and SVM, we developed hierarchical methods with RF. Previous hierarchical methods used RF as a feature selection method because it is well suited to a data set of high dimensionality with varied data types, to identify a subset of features to feed into methods such as LR and SVM.[22] RF is well known to use out-of-bag estimates and an internal bootstrap to help reduce and select only predictive variables and avoid overfitting, similar to AdaBoost.[32] RF is well known to be able to take varied data types and high-dimensional data and reduce to a usable subset, which we also verify through our bootstrapped cross-validation (through the use of a derivation and validation set).[33] However, rather than to use the list of variables supplied by RF as the inputs to SVM or LR, the Tele-HF data set allowed us to use the probability predicted by RF as the inputs to SVM and to LR to create 2 new hierarchical methods.

The Tele-HF data set has comprehensive outcomes information for all patients. We leveraged this by designing 2 prediction models using all of the aforementioned methods in a hierarchical manner. We used the RF algorithm as a base for this technique. The RF model, trained on all of the available features, produced a probability for many readmission events (eg, 0–12 events in this data set). These probabilities were then given to LR and SVM as inputs from which to build models. A detailed discussion of the method by which RF and SVM as well as RF and LR are combined together is in the Data Supplement.

Analytic Approach

For each of the predictive techniques above, we iterated the analyses illustrated in Figure 1 100×. To construct the derivation and validation data sets, we split the cohort into 2 equally sized groups, ensuring equal percentages of readmitted patients in each group. To account for a significant difference in numbers of patients who were readmitted and not readmitted in each group, we weighted the ML algorithms. The weight selected for the readmitted patients was the ratio of not-readmitted patients to readmitted patients in the derivation set. The testing and selecting of appropriate weights are further detailed in the Data Supplement.

Once the derivation and validation sets were created, a traditional LR model was trained. We used SAS and the 5 most important variables as identified previously[10]: blood urea nitrogen, glomerular filtration rate, sex, waist:hip ratio, and history of ischemic cardiomyopathy. The remaining techniques were created using the full raw data of 472 inputs for each patient and were trained on the different readmission outcome labels.

We trained models to predict either the 30-day readmission or 180-readmission outcome. Although we supplied the same input data to each method, we varied the outcome information provided. Because the 180-day readmission set contains more outcomes information, including the total number of readmission events during the trial, it is possible that it might be easier to predict 180-day readmission, given the propensity of some patients to be readmitted multiple times. To provide a range of modeling for the final binary prediction (readmitted/not readmitted), we ran 5 distinct training methods based on the labels provided to the algorithm. For 30-day readmission prediction, we first generated 3 different training models with the same baseline data, but 3 different outcomes: 30-day binary outcomes, 180-day binary outcomes, and 180-day counts of readmissions. We then used the predictive models created by these 3 training methods to predict 30-day readmission outcomes in the validation cohort. Similarly, to test 180-day readmission, we generated 2 different training methods with same baseline data but different outcomes, namely, 180-day binary outcomes and 180-day counts of readmissions. A detailed description of how these methods vary from each other is available in the Data Supplement.

We ran the models generated on the validation set and calculated the area under the receiver operating characteristics curve (C statistic), which provided a measure of model discrimination. The analysis was run 100× in order to provide robustness over a potentially poor random split of patients and to generate a mean C statistic with a 95% confidence interval (CI). We also evaluated the risk-stratification abilities of each method. The probabilities of readmission generated over the 100 iterations were then sorted into deciles. Finally, we calculated the observed readmission rate for each decile to determine the predictive range of the algorithms.

These models should be used prospectively, to provide clinicians with decision-making points. For each iteration, we calculated the positive predictive value (PPV), sensitivity, specificity, and f-score, a common measurement in ML,[25] which is calculated as

$$\text{f-score} = 2 * \frac{\text{PPV} * \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}}$$

The f-score provides a balance between PPV and sensitivity, similar to a C statistic, but at designated thresholds on the ROC curve. The data-driven decision threshold for prospective modeling measurements was that which maximized the f-score.

Furthermore, to test whether a narrowed focus of prediction could improve discrimination, we conducted the same analyses using heart failure–only readmission instead of all-cause readmission.

All ML analyses were developed in R. The list of R packages used in this study is included in the Data Supplement. The Human Investigation Committee at the Yale School of Medicine approved the study protocol.

Results

Baseline characteristics of patients in 30- and 180-day analytic samples are detailed in Table 1. In both analytic samples, the proportion of women and blacks was ≈40%. Ninety percent of patients had New York Heart Association class II or III heart failure, with ≈70% having left ventricular ejection fraction of <40%. A history of hypertension (77%) and diabetes mellitus (45%) was common. The 30-day all-cause readmission rate was 17.1%, whereas the 180-day all-cause readmission rate was 48.9%. No sex-specific or race-specific differences were found.

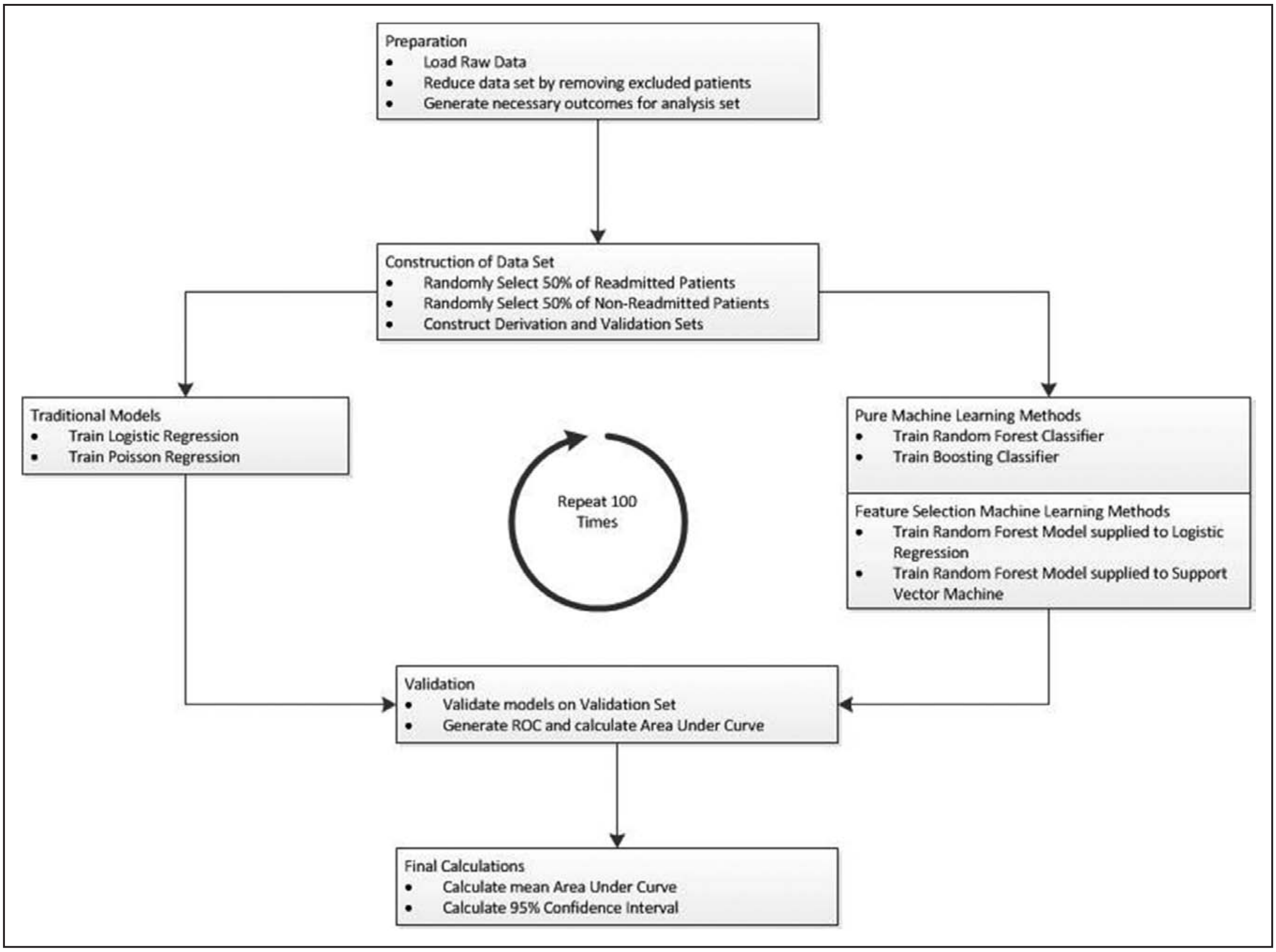

Preparation
• Load Raw Data
• Reduce data set by removing excluded patients
• Generate necessary outcomes for analysis set
Construction of Data Set
• Randomly Select 50% of Readmitted Patients
• Randomly Select 50% of Non-Readmitted Patients
• Construct Derivation and Validation Sets
Traditional Models
• Train Logistic Regression
• Train Poisson Regression
Repeat 100 Times
Pure Machine Learning Methods
• Train Random Forest Classifier
• Train Boosting Classifier
Feature Selection Machine Learning Methods
• Train Random Forest Model supplied to Logistic Regression
• Train Random Forest Model supplied to Support Vector Machine
Validation
• Validate models on Validation Set
• Generate ROC and calculate Area Under Curve
Final Calculations
• Calculate mean Area Under Curve
• Calculate 95% Confidence Interval

Figure 1. Statistical analysis flow. The data are split into derivation and validation sets. These sets are passed to each algorithm for comparison.

Thirty-Day All-Cause Model Discrimination

The discrimination of the different predictive models for 30-day all-cause readmission represented by C statistic values is illustrated in Figure 2A. LR had a low C statistic of 0.533 (95% CI, 0.527–0.538). Boosting on the input data had the highest C statistic (0.601; 95% CI, 0.594–0.607) in 30-day binary outcome with 30-day training case. Boosting also had the highest C statistic for the 30-day binary outcome with 180-day binary training (0.613; 95% CI, 0.607–0.618). For the 30-day outcomes with 180-day counts training, the RF technique had the highest C statistic (0.628; 95% CI, 0.624–0.633).

One Hundred Eighty–Day All-Cause Model Discrimination

The C statistic values for different models in predicting the 180-day readmission are illustrated in Figure 2B. LR again showed a low C statistic (0.574; 95% CI, 0.571–0.578) for the 180-day binary case. The RF into SVM hierarchical method had the highest achieved C statistic across all methods (0.654; 95% CI, 0.650–0.657) in 180-day binary outcome and 180-day count case (0.649; 95% CI, 0.648–0.651).

Thirty-Day All-Cause Model Predictive Range

Deciles of the probabilities/responses generated by the algorithms are plotted against the observed rate of readmission in each decile in Figure 3. For each training version of the method, we plotted the values from the method and training version with the highest C statistic (ie, best RF from the three 30-day training scenarios). In Figure 2A, the RF has the same mean C statistic and similar CI as the RF into LR hierarchical model. In Figure 3, for RF, the readmission rates were markedly higher at the tenth decile of probabilities (26.2%) than the first decile of probabilities (7.8%) showing a higher predictive range than LR (14.2% and 16.4%, respectively), whereas for the RF into LR hierarchical model, the actual readmission rates varied only from 10.6% to 19.0% from the first to the tenth decile, despite the same mean C statistic for RF and RF into LR.

One Hundred Eighty–Day All-Cause Model Predictive Range

When the predicted probabilities of each algorithm were separated into deciles and plotted against the real readmission rate per decile, RF into SVM had the largest difference in readmission rate from the first decile (31.0%) to the tenth decile

Table 1. Patient Characteristics

	30 d		180 d	
Characteristic	All, n (%)	Readmission, n (%)	All, n (%)	Readmission, n (%)
n	1004 (100.0)	149 (14.8)	977 (100.0)	478 (48.9)
Median age, y (SD)	62 (15.7)	62 (15.9)	62 (15.7)	62 (15.4)
Women	415 (41.3)	58 (38.9)	403 (41.2)	193 (40.4)
Race				
White	507 (50.5)	84 (56.4)	491 (50.3)	253 (52.9)
Black	393 (39.1)	55 (36.9)	385 (39.4)	188 (39.3)
Other	104 (10.4)	10 (6.7)	101 (10.3)	37 (7.8)
New York Heart Association				
Class I	56 (5.6)	7 (4.7)	54 (5.5)	31 (6.5)
Class II	515 (51.3)	85 (57.0)	500 (51.2)	256 (53.6)
Class III	355 (35.4)	52 (35.0)	347 (35.5)	160 (33.4)
Class IV	58 (5.8)	2 (1.3)	57 (5.8)	20 (4.2)
Missing	20 (1.9)	3 (2.0)	19 (2.0)	11 (2.3)
Medical history				
LVEF % <40	687 (68.4)	99 (66.4)	668 (68.4)	317 (66.3)
Hypertension	771 (76.8)	116 (77.9)	752 (77.0)	376 (78.7)
Diabetes mellitus	450 (44.8)	57 (38.3)	439 (44.9)	200 (41.8)
Myocardial infarction	257 (25.6)	47 (31.5)	250 (25.6)	131 (27.4)
Stroke	96 (9.6)	15 (10.1)	92 (9.4)	44 (9.2)
Ischemic cardiomyopathy	235 (23.4)	44 (29.5)	228 (23.3)	127 (0.27)
Clinical values, mean/SD				
Albumin	3.32 (0.53)	3.25 (0.61)	3.31 (0.53)	3.33 (0.55)
Blood urea nitrogen	25.2 (17.8)	28.7 (20.3)	26.3 (16.8)	29.1 (17.5)
Creatinine	1.40 (0.77)	1.53 (0.80)	1.45 (0.72)	1.55 (0.80)
Hemoglobin	12.3 (1.94)	12.0 (1.80)	12.4 (1.94)	12.1 (1.87)
Glomerular filtration rate	58.5 (27.4)	52.8 (24.6)	58.8 (27.4)	54.4 (27.2)
Potassium	4.08 (0.57)	4.19 (0.58)	4.08 (0.57)	4.09 (0.56)

LVEF indicates left ventricular ejection fraction.
*All values are mean (SD) unless noted.

(69.4%), whereas LR had a readmission rate of 40.5% in the first decile and 60.1% in the tenth decile (Figure 4).

Readmission Because Of Heart Failure Discrimination

As illustrated in Figure 5A and 5B, for readmissions because of heart failure, the LR model again had a low C statistic for the 30-day binary case (0.543; 95% CI, 0.536–0.550) and the 180-day binary case (0.566; 95% CI, 0.562–0.570). Boosting had the best C statistic for the 30-day binary-only case (0.615; 95% CI, 0.607–0.622) and for the 30-day with 180-day binary case training (0.678; 95% CI, 0.670–0.687). The highest C statistic for other prediction cases were RF for the 30-day with 180-day counts case training (0.669; 95% CI, 0.661–0.676); RF into SVM for the 180-day binary-only case (0.657; 95% CI, 0.652–0.661); and RF into SVM for the 180-day counts case (0.651; 95% CI, 0.646–0.656).

Readmission Because Of Heart Failure Predictive Range

When the deciles of risk prediction were plotted against the observed readmission rate, RF and boosting each had the biggest differences between the first and tenth deciles of risk (1.8–11.9% and 1.4–12.2%, respectively). Although the hierarchical methods of RF into LR and RF into SVM had similar C statistics (0.654; 95% CI, 0.645–0.663 and 0.656; 95% CI, 0.648–0.664, respectively) in Figure 6, only RF into LR clearly identifies low- and high-risk groups. The risk stratification for the 180-day case because of heart failure follows closely with C statistics, similar to that for the 180-day all-cause case (Figure 7).

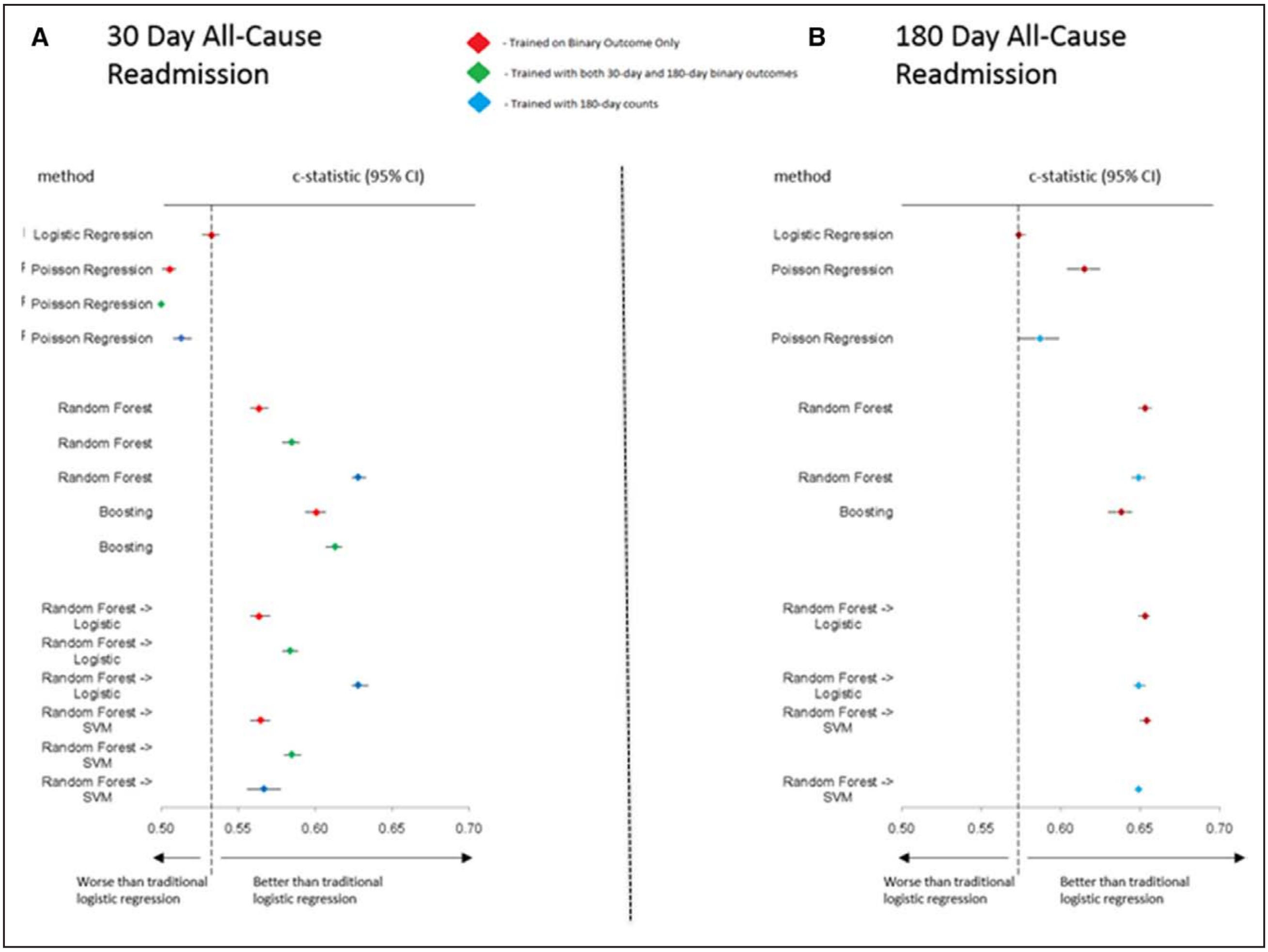


Figure 2. Forest plots for C statistics and 95% confidence intervals (CIs) for each method with respect to prediction of 30- and 180-d all-cause readmission (**A** and **B**). Diamond represents C statistic and the line represents the 95% CI. Color coding for model training: red: trained in the 30-d binary case only; green: trained with 30- and 180-d binary outcome; blue: trained with the 180-d counts case. SVM indicates support vector machine.

Improvements Over LR

Table 2 shows the percentage improvement in prediction of the best-trained model for each ML technique over LR for each outcome predicted. RF achieved a 17.8% improvement in discrimination over LR, the best 30-day all-cause readmission improvement. Similarly, the hierarchical method of RF into SVM achieved a 13.9% improvement for 180-day all-cause readmission; boosting achieved a 24.9% improvement for 30-day readmission because of heart failure; and the hierarchical method of RF into SVM achieved a 16.1% improvement over LR for 180-day readmission because of heart failure.

Prediction Results

Table 3 shows the prospective prediction results of the best model for 30-day all-cause readmission, 180-day all-cause readmission, 30-day readmission because of heart failure, and 180-day readmission because of heart failure. Thirty-day all-cause readmission had a PPV of 0.22 (95% CI, 0.21–0.23), sensitivity of 0.61 (0.59–0.64), and a specificity of 0.61 (0.58–0.63) at a maximal f-score of 0.32 (0.31–0.32); 180-day all-cause readmission had a PPV of 0.51 (0.51–0.52), a sensitivity of 0.92 (0.91–0.93), and a specificity of 0.18 (0.16-0.21) at a maximal f-score of 0.66 (0.65–0.66); 30-day readmission because of heart failure had a PPV of 0.15 (0.13–0.16), a sensitivity of 0.45 (0.41–0.48), and a specificity of 0.79 (0.76–0.82) at a maximal f-score of 0.20 (0.19–0.21); 180-day readmission because of heart failure had a PPV of 0.51 (0.50–0.51), a sensitivity of 0.94 (0.93–0.95), and a specificity of 0.15 (0.13–0.17) at a maximal f-score of 0.66 (0.65–0.66). The 30-day predictions, in general, were better at identifying negative cases, whereas the 180-day predictions were better able to correctly identify positive cases.

Discussion

In our study to model risk of readmissions in a sample of patients hospitalized with heart failure, the ML analytic techniques we applied improved discrimination modestly compared with the traditional method of LR. These ML models also provided better identification of groups at low and high risk for readmission by increasing predictive range compared with LR. These results are consistent with our hypothesis that ML methods can leverage complex higher-level interactions among a multitude of variables to improve discrimination and

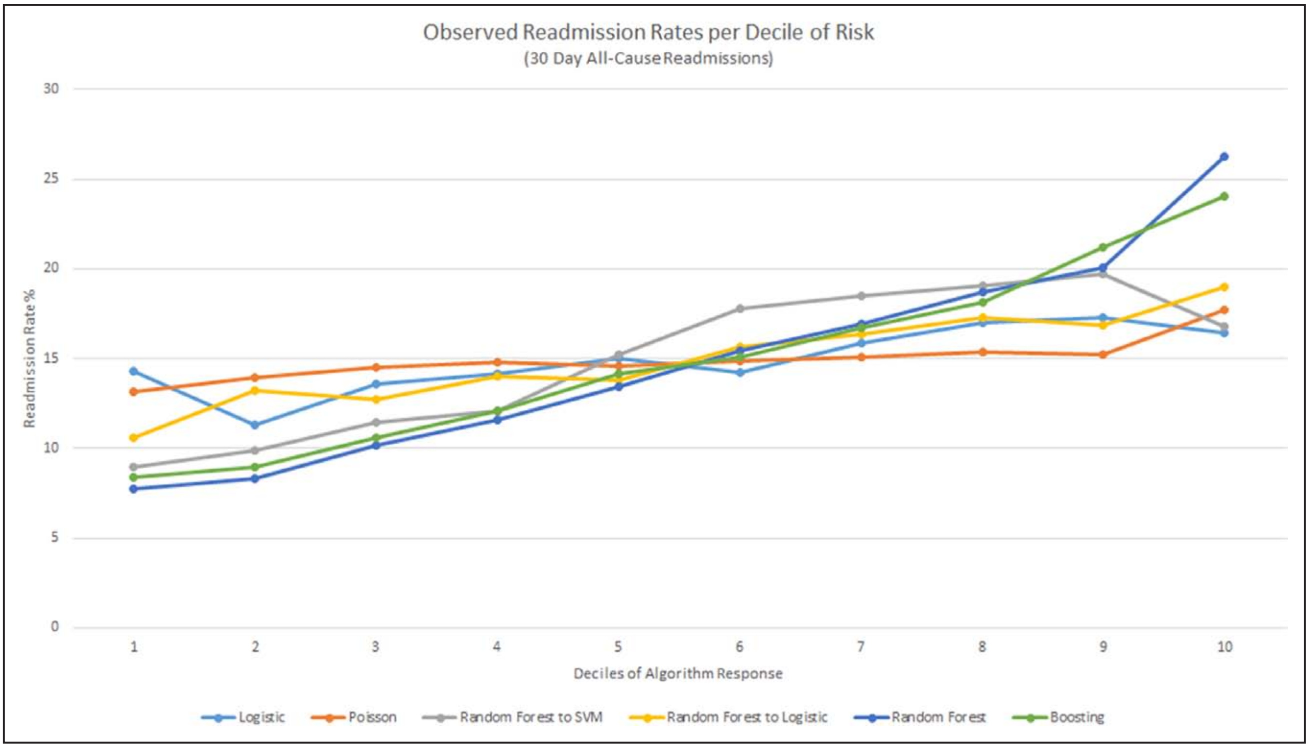


Figure 3. Deciles of algorithm risk versus readmission rates (%) for the best 30-d all-cause models for each method. The *y* axis presents the observed readmission rates in each decile, the *x* axis the ordered deciles of risk predicted. SVM indicates support vector machine.

predictive range with respect to heart failure outcomes compared with traditional linear models. We used readmissions among patients with heart failure as a test case to examine the advantage of ML algorithms over traditional statistical analysis to see whether such an approach could be applied to other outcomes prediction and predictive range problems.

Models developed from ML algorithms to improve discrimination and predictive range can be evaluated using

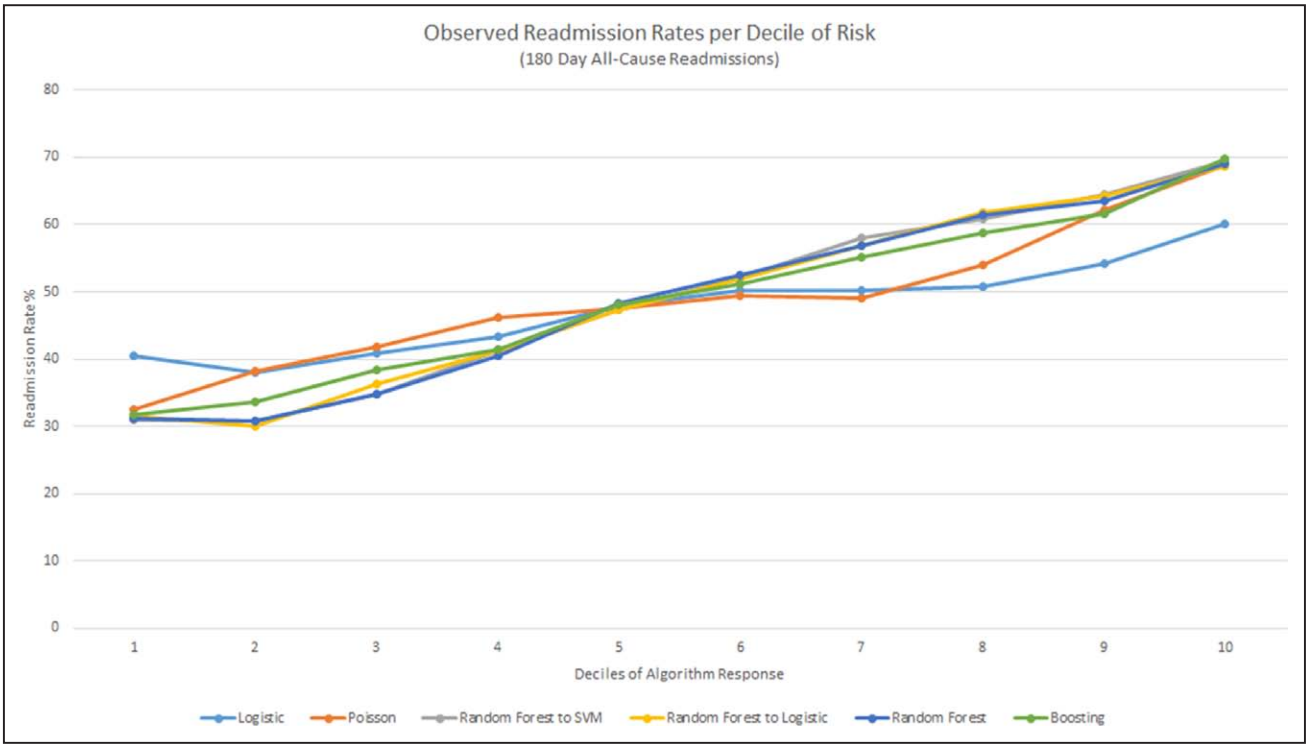


Figure 4. Deciles of algorithm risk versus readmission rates (%) for the best 180-d all-cause models for each method. The *y* axis presents the observed readmission rates in each decile, the *x* axis the ordered deciles of risk predicted. SVM indicates support vector machine.

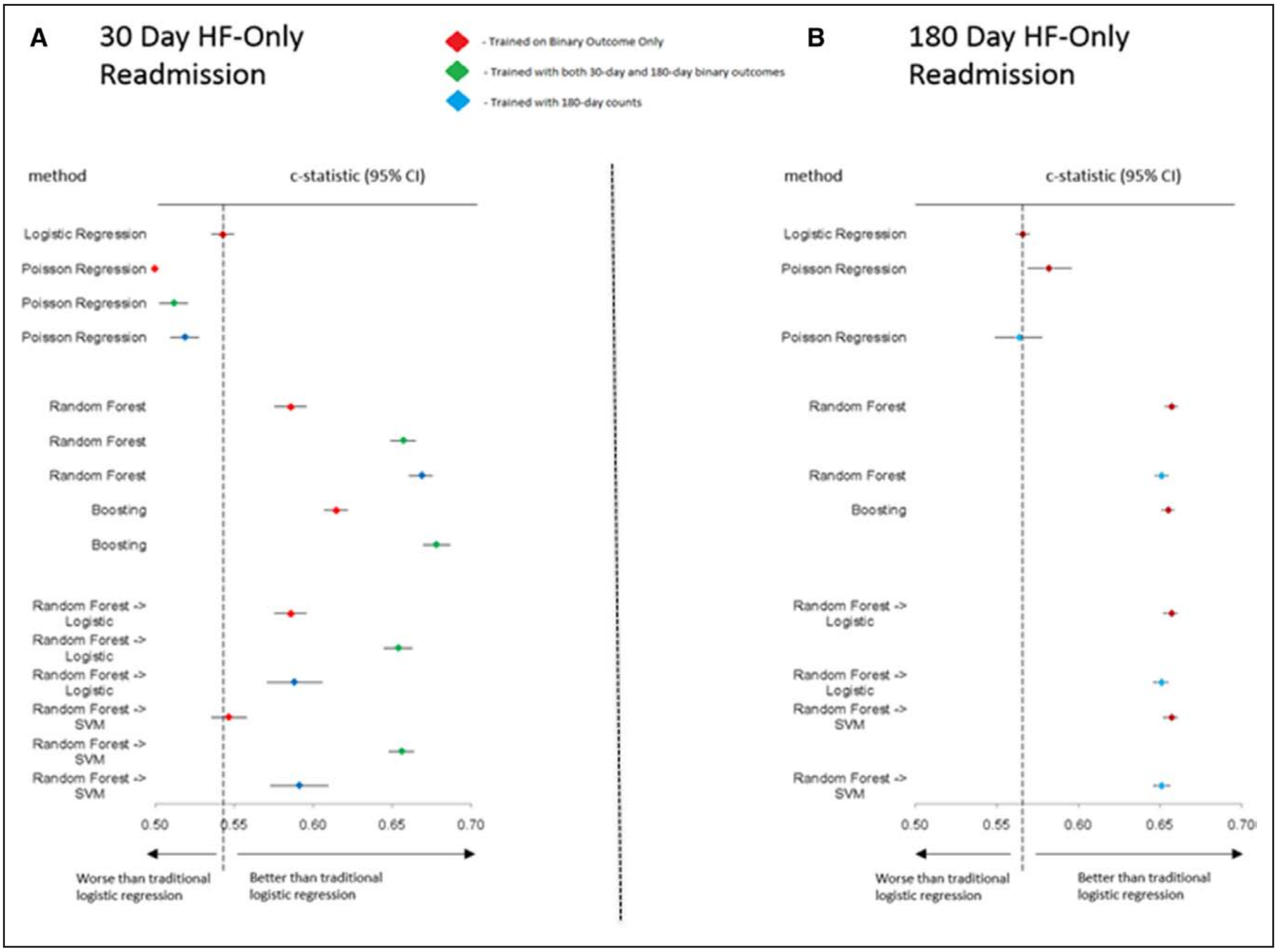


Figure 5. Forest plots for C statistics and 95% confidence intervals (CIs) for each method with respect to prediction of 30- and 180-d heart failure (HF) readmission (**A** and **B**). Diamond represents C statistic, and the line represents the 95% CI. Color coding for model training: red: trained in the 30-d binary case only; green: trained with 30- and 180-d binary outcome; blue: trained with the 180-d counts case. SVM indicates support vector machine.

several measures or methods. The modest improvement in discrimination represented by C statistics across the various ML methods, while promising, is ultimately insufficient. Furthermore, the models selected, at a data-driven threshold that optimizes the f-score, optimized the correct classification of negative cases but had a large number of false positives, as seen in the low PPV and f-score. This threshold decision would be varied if further cost information was inputted to the model or the selected threshold optimized on another factor, such as reducing false alarms. We found that although discriminatory power was similar across certain methods, the range in observed events among deciles of predicted risk varied greatly. Analyzing multiple aspects of models provides a stronger understanding of those models. With regard to 30-day all-cause readmission case, RF better differentiates low- and high-risk groups compared with all other methods including the hierarchical RF into LR model. Despite similar discrimination (mean C statistic) between the 2 methods, RF actually has a better predictive range. In other words, RF better discriminates patients in low- and high-risk strata, whereas hierarchical RF into LR better discriminates patients in the middle deciles of risk. This suggests that a combination of various approaches, with possible deployment of different methods at different strata of risk, can improve discrimination of the overall model and may improve the predictive range.

There have been some suggestions in the literature about how to improve predictive range of risk strata and overall discrimination. Wiens et al.[26] showed that use of time-based information can improve their predictive range and overall discrimination when predicting the risk of *Clostridium difficile* infection in hospitalized patients. In an earlier study examining heart failure mortality prediction using cytokine measurements, use of the baseline cytokine measurements alone did not significantly improve discrimination.[34] However, the addition of serial cytokine measurements over follow-up time to the predictive model improved discrimination for 1-year mortality.[34] Thus, perhaps leveraging serially collected telemonitoring data on daily health and symptoms of patients could improve discrimination and predictive range of our predictions.

The general experience with LR is that readmissions after index hospitalizations for heart failure are more difficult to predict than mortality outcomes. In a meta-analysis, Ouwerkerk et al[12] analyzed 117 heart failure outcome prediction models

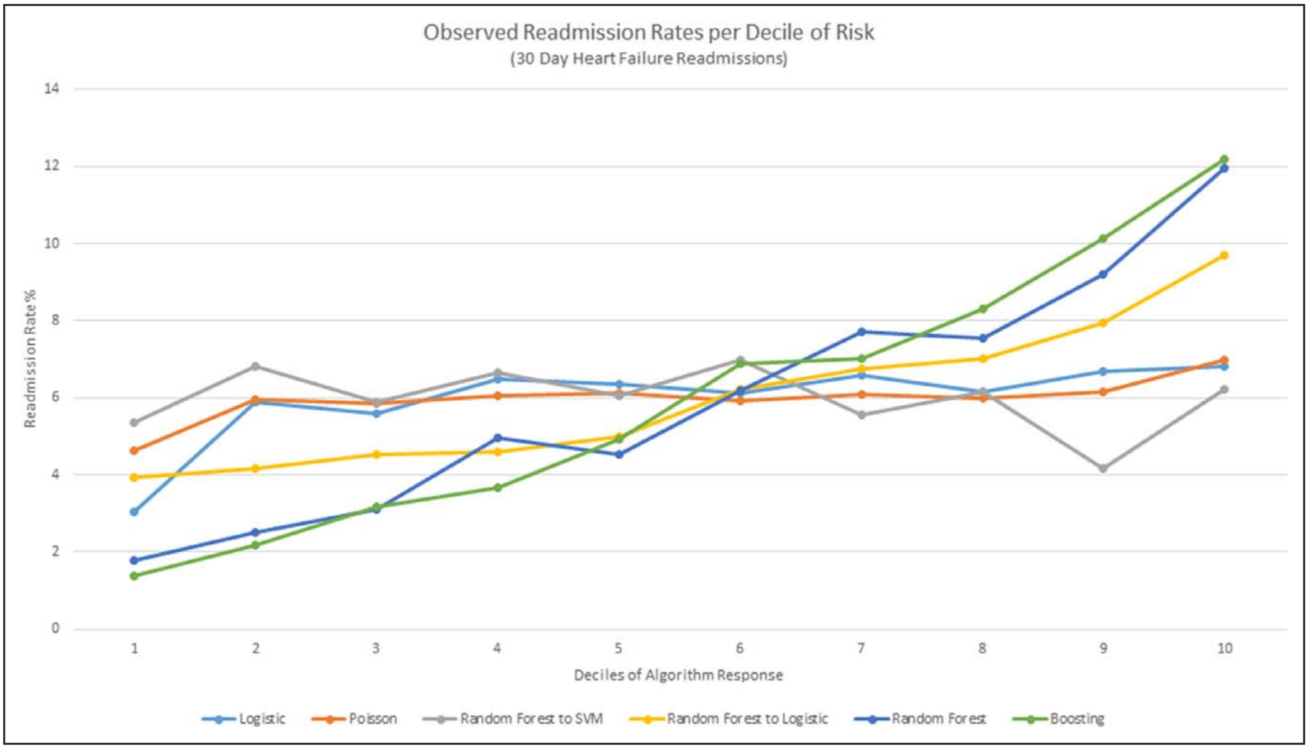


Figure 6. Deciles of algorithm risk vs readmission rates (%) for the best 30-d heart failure–only models for each method. The *y*-axis presents the observed readmission rates in each decile, the *x*-axis the ordered deciles of risk predicted. SVM indicates support vector machine.

with 249 variables; the average C statistic to predict mortality was 0.71 and hospitalization was 0.63. Tele-HF study data were collected to identify and track heart failure–related symptoms, and we find that narrowing our prediction to heart failure–only readmissions improves the discrimination and predictive range of each method.

Ultimately, despite the increase in a wide array of data, from quality of life questionnaires to hospital examinations,

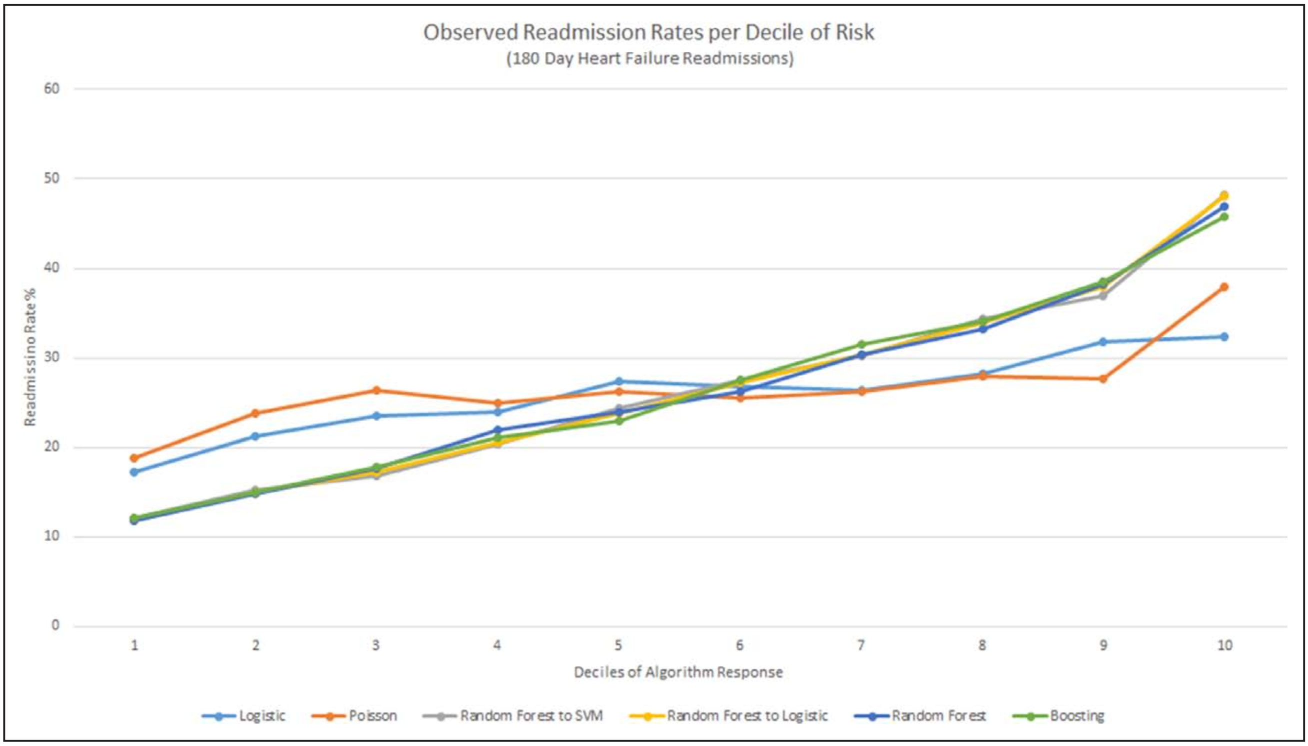


Figure 7. Deciles of algorithm risk versus readmission rates (%) for the best 180-d heart failure–only models for each method. The *y*-axis presents the observed readmission rates in each decile, the *x*-axis the ordered deciles of risk predicted. SVM indicates support vector machine.

Table 2. Percent Improvement of Machine Learning Methods Over Baseline, Traditional Logistic Regression

	Poisson Regression, %	Random Forest, %	Boosting, %	Random Forest to Logistic Regression, %	Random Forest to Support Vector Machine, %
30-d all-cause readmission prediction	–3.75	17.8	15.0	17.8	9.76
180-d all-cause readmission prediction	7.14	13.8	11.0	13.8	13.9
30-d heart failure–only readmission prediction	–5.71	23.2	24.9	20.4	20.8
180-d heart failure–only readmission prediction	2.83	16.1	15.7	16.1	16.1

as well as the increase in complex methods to perform predictions, questions arise from the modest C statistics. The incremental gains over the LR models indicate the power of the ML methods. However, a deeper analysis of why, even with data from a multitude of domains, the models remain at a consistent level with the previous work raises the question of what other data might be collected that would provide predictive value, and whether it is currently being collected or not. It seems that a deeper study into other possible important predictors of readmission might need to be performed to better understand why current administrative claims and clinical studies models have had varied, but moderate, success.

Limitations

The discrimination and range of prediction are likely affected by several limitations of the Tele-HF data set. First, the number of patients is small in comparison to the number of variables measured per patient. As a result, while rich in terms of patient features, the small number of patients makes the data set vulnerable to overfitting. Furthermore, many variables included in models with stronger C statistics from previous studies are not present in Tele-HF,[9,27] which could affect performance. Time-to-event analysis conducted on the Tele-HF data set produced a C statistic of 0.65.[10] This finding suggests there are measurable domains of variables not analyzed in this study that might be more predictive in all patients with heart failure. However, with the moderate-at-best C statistics, it is possible that the best set of features in predicting risk in patients with heart failure has not yet been collected.

The interpretation of the results can be varied depending on the particular data set and purpose of modeling. For example, the prediction threshold selected at the maximal f-score may not be desirable for a particular intervention designed when considering a model, such as when used in a prospective fashion with the false alarm rate in mind. Furthermore, the numeric ranges of the probabilities generated by the algorithms do not necessarily match, and thus the thresholds considered across each method vary. Although the output of the algorithm provides a predicted probability of readmission, the value in this is not the number provided but where this lies in the predictive range of the method and, thus, whether it can be classified as low or high risk, where the observed rates are considered. This is a result of each method basing its distance (and thus probability) of observations in the model on different metrics, resulting in a different calculation of a final probability. If these probabilities would be desirable for direct use, the values could be scaled to match comparable intervals across methods.

The trial arm of Tele-HF may provide further insights into the value of time-to-event analysis by permitting inclusion of the daily telemonitoring responses by patients in the intervention arm as a time-series signal to be processed for time-dependent risk models. To determine the accuracy of additional methods, a direct comparison in the predictions of each individual patient (both risk and outcome) needs to be made between methods rather than net reclassification index work, which has not been shown to be effective in many cases at understanding and improving model discrimination performance.[35,36] Finally, to address the viability of the model and the clinical predictors it is built on, an external validation cohort should be used. However, as the clinical predictors collected in studies are often varied, it becomes important to distinguish unique predictors in internal validation, as well as understand the impact each predictor has on the models for better interpretation. As this is a post hoc analysis of a randomized controlled trial, the generalizability of such models is a limitation. Nevertheless, the goal of this study was to compare methods, not to introduce a new model. Therefore, there is little reason to think that a population-based sample would yield a different result. Although the data from such a trial span a diverse set of hospitals and patients, and is well curated, once the understanding of such strengths and weaknesses of ML methods is understood, they should be extended

Table 3. Prospective Prediction Results (Mean and 95% Confidence Intervals)

	Positive Predictive Value	Sensitivity	Specificity	F-Score
30-d all-cause readmission	0.22 (0.21–0.23)	0.61 (0.59–0.64)	0.61 (0.58–0.63)	0.32 (0.31–0.32)
180-d all-cause readmission	0.51 (0.51–0.52)	0.92 (0.91–0.93)	0.18 (0.16-0.21)	0.66 (0.65–0.66)
30-d heart failure–only readmission	0.15 (0.13–0.16)	0.45 (0.41–0.48)	0.79 (0.76–0.82)	0.20 (0.19–0.21)
180-d heart failure–only readmission	0.51 (0.50–0.51)	0.94 (0.93–0.95)	0.15 (0.13–0.17)	0.66 (0.65–0.66)

to data derived from other sources as well, including electronic health records, which will likely affect results because of a consideration of varied data availability and a more heterogeneous patient population.

Conclusions

The results of our study support the hypothesis that ML methods, with the ability to leverage all available data and their complex relationships, can improve both discrimination and range of prediction over traditional statistical techniques. The performance of the ML methods varies in complex ways, including discrimination and predictive range. Future work needs to focus on further improvement of predictive ability through advanced methods and more discerning data, so as to facilitate better targeting of interventions to subgroups of patients at highest risk for adverse outcomes.

Sources of Funding

Dr Dharmarajan was supported by grant K23AG048331 from the National Institute on Aging, Bethesda, MD, and the American Federation for Aging Research through the Paul B. Beeson Career Development Award Program, New York, NY. Dr Krumholz was supported by grant U01 HL105270 (Center for Cardiovascular Outcomes Research at Yale University) from the NHLBI in Bethesda, MD. Dr Manhapra was supported by the Research in Addiction Medicine Scholars (RAMS) Program, R25DA033211 from the NIDA, Rockville, MD. The data used in this work were from the Tele-HF trial (Telemonitoring to Improve Heart Failure Outcomes; NCT00303212). The funders had no role in the design and conduct of the study; in collection, analysis, and interpretation of data; and in preparation, review, or approval of article.

Disclosures

Drs Dharmarajan and Krumholz work under contract with the Centers for Medicare & Medicaid Services to develop and maintain performance measures. Dr Dharmarajan serves on a scientific advisory board for Clover Health, Inc. Dr. Krumholz is chair of a cardiac scientific advisory board for UnitedHealth, is the recipient of research agreements from Medtronic and Johnson and Johnson (Janssen), to develop methods of clinical trial data sharing, and is the founder of Hugo, a personal health information platform. The other authors report no conflicts.

References

1. Kocher RP, Adashi EY. Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. *JAMA*. 2011;306:1794–1795. doi: 10.1001/jama.2011.1561.
2. Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, Barreto-Filho JA, Kim N, Bernheim SM, Suter LG, Drye EE, Krumholz HM. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*. 2013;309:355–363. doi: 10.1001/jama.2012.216476.
3. Ross JS, Chen J, Lin Z, Bueno H, Curtis JP, Keenan PS, Normand SL, Schreiner G, Spertus JA, Vidán MT, Wang Y, Wang Y, Krumholz HM. Recent national trends in readmission rates after heart failure hospitalization. *Circ Heart Fail*. 2010;3:97–103. doi: 10.1161/CIRCHEARTFAILURE.109.885210.
4. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306:1688–1698. doi: 10.1001/jama.2011.1515.
5. Dharmarajan K, Krumholz HM. Strategies to reduce 30-day readmissions in older patients hospitalized with heart failure and acute myocardial infarction. *Curr Geriatr Rep*. 2014;3:306–315. doi: 10.1007/s13670-014-0103-8.
6. Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, Krumholz HM. Statistical models and patient predictors of readmission for heart failure: a systematic review. *Arch Intern Med*. 2008;168:1371–1386. doi: 10.1001/archinte.168.13.1371.
7. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail*. 2014;2:440–446. doi: 10.1016/j.jchf.2014.04.008.
8. Lupón J, Vila J, Bayes-Genis A. Risk prediction tools in patients with heart failure. *JACC Heart Fail*. 2015;3:267. doi: 10.1016/j.jchf.2014.10.010.
9. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, Reed WG, Swanson TS, Ma Y, Halm EA. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care*. 2010;48:981–988. doi: 10.1097/MLR.0b013e3181ef60d9.
10. Krumholz HM, Chaudhry SI, Spertus JA, Mattera JA, Hodshon B, Herrin J. Do non-clinical factors improve prediction of readmission risk?: Results from the Tele-HF study. *JACC Heart Fail*. 2016;4:12–20. doi: 10.1016/j.jchf.2015.07.017.
11. Levy WC, Anand IS. Heart failure risk prediction models: what have we learned? *JACC Heart Fail*. 2014;2:437–439. doi: 10.1016/j.jchf.2014.05.006.
12. Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail*. 2014;2:429–436. doi: 10.1016/j.jchf.2014.04.006.
13. Weiss GM, Lockhart JW. The impact of personalization on smartphone-based activity recognition. AAAI Workshop on Activity Context Representation: Techniques and Languages. 2012.
14. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;7:3. doi: 10.1186/1471-2105-7-3.
15. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*. 2008;9:319. doi: 10.1186/1471-2105-9-319.
16. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning*. Springer, 2009.
17. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. 2009;10:213. doi: 10.1186/1471-2105-10-213.
18. Bergstra J, Casagrande N, Erhan D, Eck D, Kégl B. Aggregate features and AdaBoost for music classification. *Mach Learn*. 2006;65:473–484.
19. Khammari A, Nashashibi F, Abramson Y, Laurgeau C. Vehicle detection combining gradient analysis and AdaBoost classification. Intelligent Transportation Systems, 2005 Proceedings; IEEE. 2005:66–71.
20. Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E. Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS One*. 2014;9:e109264. doi: 10.1371/journal.pone.0109264.
21. Chen Y-W, Lin C-J. *Combining SVMs With Various Feature Selection Strategies Feature Extraction*. Springer, 2006:315–324.
22. Pal M, Foody GM. Feature selection for classification of hyperspectral data by SVM. Geoscience and Remote Sensing, IEEE Transactions on. 2010;48:2297–2307.
23. Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. Pattern Recognition, 2004 ICPR 2004 Proceedings of the 17th International Conference on. 2004;3:32–36.
24. Wang S, Zhu W, Liang Z-P. Shape deformation: SVM regression and application to medical image segmentation. Computer Vision, 2001 ICCV 2001 Proceedings Eighth IEEE International Conference on. 2001;2:209–216.
25. Mortazavi B, Pourhomayoun M, Nyamathi S, Wu B, Lee SI, Sarrafzadeh M. Multiple model recognition for near-realistic exergaming. Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on. 2015:140–148.
26. Wiens J, Horvitz E, Guttag JV. Patient risk stratification for hospital-associated c. diff as a time-series classification task. *Adv Neural Inf Process Syst*. 2012:467–475.
27. Wang L, Porter B, Maynard C, Bryson C, Sun H, Lowy E, McDonell M, Frisbee K, Nielson C, Fihn SD. Predicting risk of hospitalization or death among patients with heart failure in the Veterans Health Administration. *Am J Cardiol*. 2012;110:1342–1349. doi: 10.1016/j.amjcard.2012.06.038.
28. Coxe S, West SG, Aiken LS. The analysis of count data: a gentle introduction to poisson regression and its alternatives. *J Pers Assess*. 2009;91:121–136. doi: 10.1080/00223890802634175.
29. Chaudhry SI, Mattera JA, Curtis JP, Spertus JA, Herrin J, Lin Z, Phillips CO, Hodshon BV, Cooper LS, Krumholz HM. Telemonitoring in patients

with heart failure. *N Engl J Med*. 2010;363:2301–2309. doi: 10.1056/NEJMoa1010029.

30. Chaudhry SI, Barton B, Mattera J, Spertus J, Krumholz HM. Randomized trial of Telemonitoring to Improve Heart Failure Outcomes (Tele-HF): study design. *J Card Fail*. 2007;13:709–714. doi: 10.1016/j.cardfail.2007.06.720.
31. Dharmarajan K, Krumholz HM. Opportunities and challenges for reducing hospital revisits. *Ann Intern Med*. 2015;162:793–794. doi: 10.7326/M15-0878.
32. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
33. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett*. 2010;31:2225–2236.
34. Subramanian D, Subramanian V, Deswal A, Mann DL. New predictive models of heart failure mortality using time-series measurements and ensemble models. *Circ Heart Fail*. 2011;4:456–462. doi: 10.1161/CIRCHEARTFAILURE.110.958496.
35. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*. 2014;25:114–121. doi: 10.1097/EDE.0000000000000018.
36. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33:3405–3414. doi: 10.1002/sim.5804.

Circulation
Cardiovascular Quality and Outcomes

American Heart Association®

Analysis of Machine Learning Techniques for Heart Failure Readmissions

Bobak J. Mortazavi, Nicholas S. Downing, Emily M. Bucholz, Kumar Dharmarajan, Ajay Manhapra, Shu-Xia Li, Sahand N. Negahban and Harlan M. Krumholz

Circ Cardiovasc Qual Outcomes. 2016;9:629-640; originally published online November 8, 2016;
doi: 10.1161/CIRCOUTCOMES.116.003039
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2016 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

SUPPLEMENTAL MATERIAL

Mortazavi: Analysis of Machine Learning Techniques for Heart Failure Readmissions

Supplemental Methods

Data Set Creation

Variables

This section details the creation of the features, the alignment of particular features, and the missing features. In particular, **Table S1** shows the list of all baseline variables collected in the Tele-HF trial. Removing Patient ID, the 236 remaining variables each had a duplicate binary variable created to indicate whether that value was missing or not (e.g., AGE_MISSING would be a binary variable with "yes" if the age was missing for that given patient and "no" otherwise).

Data Alignment

The values in the variables are of three types, either binary, continuous, or categorical. However, due to the combination of multiple questionnaire types, the values assigned to the binary and categorical answers do not match. For example, one categorical value may have a numeric range from 1 to 5 where 1 is the worst answer and 5 is the best. The very next question may have five options as well but order them 0 to 4 where 4 is the worst. As a result, before any algorithms were run, all of the variables were ordered to increase as the answer improved (aligning the intensities as much as possible). The worst answer was given a value of 1 and the best answer would go up from there, 5 in most cases but up to 7 or 8 in many. Thus, the missing values would result in 0 being assigned. In this case the missing value provides no weight to the machine learning algorithm and can be considered either its own category or aligned with a truly negative response. This was chosen to integrate well with future endeavors on the telemonitoring data where the hypothesis to be tested is that missing data correlates strongly with negative responses and adverse outcomes.

Predictive Methods

Setting up the Outcome

Before we discuss the algorithms in further detail we should further outline how we arrived at the conclusion that we should weight our readmission cases heavily (equal to the proportion of not-readmitted vs. readmitted cases) versus the many other possibilities run in the 30-day prediction case (since the 180-day case had a roughly equal number of readmits to non-readmits). We ran the following iterations of dealing with class imbalance. We first applied all the algorithms with no weighting and found that the probability of being predicted as a readmit was quite low. In order to balance the sets, we tried a number of techniques. We downsampled the non-readmit cases to be equal to the number of readmit cases in the training set, but found that the number of training samples was simply too small to create an effective model. We then upsample the readmitted cases to be equal to the number of non-readmit cases. This improved some of the algorithms and not all. The final approach we took was to vary our weighting of the readmit cases versus the non-readmit cases. Comparing the results of no weighting, downsampling, and upsampling, we observed that the weights setting the two sets to roughly equal seemed to be the best (which it was). To further confirm this observation, we varied the weight across a range from no weighting to twice as strong as the proportion of non-readmits to readmits. The rest of the methods described were run for each of these cases but the final outputs presented in the paper concern only the weighted examples.

Logistic Regression (LR) Comparisons

Three different LR models were built in order to validate the strength of the model presented in the main text, as well as to validate its use as a comparative technique to the machine learning (ML) models built. The first such model, based on GLMNET[1], was originally used for LR with Lasso regularization for feature selection, but the cross-validated LR in this model deemed no variables worthy of including in the final model so it produced a C-statistic of 0.5 since the only feature in the model was the intercept.

The second such model used the 236 Tele-HF variables and computed a forward, stepwise selection based upon the likelihood ratio of each variable. Forward selection was chosen because in certain high-dimensional datasets, the results are actually considered closer to the optimal solution than a Lasso regularization as in GLMNET.[2] The method was run until no variable's likelihood had a p-value < 0.01. This technique produced a model, in 100 bootstrapped iterations, with a mean C-statistic of 0.524 with a 95% confidence interval over the 100 iterations of (0.518-0.529). Further, the model selected on average only 4.79 (95% CI: 4.5-5.1) variables. These variables selected were different in each iteration, with 80 of the 236 variables being selected across the 100 iterations. The variables selected are listed in **Table S2**. Note that the frequently-selected variables do not match those from Krumholz et al., indicating their valuation of variable importance results in the selection of more appropriate predictors**.**

The third such model used the full 472 variables created for this paper, including the dummy missing variables. Features were forward-selected similarly, and the method produced a C-statistic of 0.518 (0.512-0.524) with on average 5.53 (5.2-5.9) variables. Incidentally, this method selected 85 of the 472 variables, none of which were the missing dummy variables. These variables served only to affect the likelihood calculations and selections of the variables considered in the second method above, and increasing the number of variables beyond 5 actually lowered the C-statistic. As a result, the method chosen in the main text, based upon a prior study that comprehensively reviewed and evaluated the predictive capabilities of each variable in the Tele-HF dataset, produces the best and fairest comparison technique for the work.

Inputs and Outputs to Each Method

This section will cover in greater detail the machine learning algorithms considered and how they were coded in R for replication. SAS was used (proc logistic) to model the logistic regression as explained in the main text. The remainder of the techniques were coded in R.

The first ML technique, Poisson Regression[1], uses the input data as well as the total number of readmissions to create a predictive model based upon the propensity to be readmitted.

This technique is classically used to predict a range of counts (e.g., the number of readmission events) it can also be used for comparing the binary case of readmitted/not readmitted, has a built in feature ordering and selection technique, and outputs the variables associated with the predictive model along with a propensity value for prediction. This value is given to the pROC package along with the ground truth labels to determine the ROC curve and area under the curve estimate.[3]

The second technique, Random Forests[4] (RF), uses a series of decision trees on the input data to predict a final outcome by considering the result of each decision tree. The decision trees can be trained to output a binary decision or a prediction on the range of readmissions. Further, the ordering of features in the trees gives a selection of the most important features used for prediction. Finally, RF can be trained using the number of readmissions or the binary label of readmitted/not readmitted and can output probabilities of a binary prediction or a multiclass prediction (e.g., probability of each particular number of readmissions). For each of the test cases we had RF output the probabilities of being within each class. For the binary readmission case, the probability of being in class 1 (readmitted) was used for pROC calculations. For the counts of readmissions, we are given a matrix of probabilities where each column indicates each count, namely, 0 for no readmissions, 1 for 1 readmission, 2 for 2 readmissions and so on. We tested a combination of factors to add probabilities for a final 0 or 1 prediction. We tried the final 0 prediction probability being only the column associated with class 0. We then added the probability of being in class 1 with class 0 and called this a probability of not being readmitted, and so forth through all of the probabilities. When being fed into LR or support vector machines (SVM), it was these matrix of probabilities that were supplied as inputs. The results presented in the paper indicate the best form of this approach, where class 0 was considered no readmissions and the probabilities of all the other classes were added together to form the probability of being readmitted.

The third technique, that also contains a form of feature ranking, is Boosting. Boosting attempts to take a series of weak classifiers (e.g., tree classifiers based upon a single feature) that only classify pieces of the data well, and re-weigh them to develop an overall strong classifier. This iterative technique then builds a strong classifier as a weighted linear combination of the results of these weak classifiers, to output a binary outcome. We used the ADA package[5] to test this method. This package has the flexibility of providing two loss functions, exponential and logistic, as well as the different kind of boosting techniques, discrete (also known as AdaBoost), real (for RealBoost), and gentle (for GentleBoost). The results presented in the paper were a summary of the best form of Boosting calculated by the algorithms.

The final ML technique considered is the SVM, implemented in R by the package e1071, of which the SVM implementation is known as LibSVM[6]. An SVM is a supervised learning model that leverages higher dimensional spaces to attempt to determine separation between the classes being trained. Unlike the previous methods, however, a feature selection algorithm must be run prior to the SVM to select the most relevant features. Such selection algorithms can be take many forms, and in this work, will be provided by RF, where the output from the RF models (the matrix of probabilities) will serve as the inputs to the SVM. Both linear and radial basis function kernels (RBF) were used but in each case the RBF algorithm outperformed the linear kernel. In some cases, the linear kernel was unable to converge on a solution in the given number of iterations.

Again, in all cases, instead of looking at a final response output, we looked at the probabilities of being readmitted generated by the algorithm, then supplied that to the pROC package to vary thresholds of readmitted/not readmitted and generate an ROC curve for us.

Using RF with Other Methods

RF is a method that is well-suited to a dataset of high dimensionality with a number of mixed types.[7] RFs build decision trees where each node is split by a single chosen variable. This variable is selected by using out-of-bag estimates and bootstrapping to measure error, correlation to other variables, and strength of prediction, to pick the best variables as well as ensure the model does not overfit.[7] For this reason, RFs are often used in situations with a high-dimensional, varied dataset, to serve as a feature selection technique for other models as well as its own model.[8-10] For this reason, this method is often used to select features, with the top importance features used to train methods such as SVM.

This work, similarly, leverages the ability of the RF method to evaluate a large set of variables and develop a predictive model without overfitting. However, rather than taking the selected variables, which might be of different types (e.g. continuous and categorical), RF can produce the probability of multiple events. For example, rather than have a binary yes/no prediction of readmission, RF can create a regression for the number of readmissions (e.g. 0-12) and provide a probability of each readmission count. These probabilities (13 of them in this case), are then provided as inputs to SVM and LR, methods that are at risk of overfitting if all the variables were to be provided to them. Thus, the hierarchical models created avoid overfitting by leveraging RFs, which avoid overfitting by using an internal bootstrapping and out-of-bag errors to ensure this. Further, the 100 bootstrapped iterations and the accuracy produced help verify that this is the case experimentally.

Creating Deciles of Risk

Similar to the ROC creation, each iteration of responses was also split into deciles using the R quantile function. Each iteration the boundary values from the predicted responses are taken for the deciles and a mean boundary value plus 95% confidence intervals around those boundaries are calculated. The total set of responses are also then combined and split into deciles to show the fraction of times correct across the cross-validated samples. We verified that the decile boundaries, created by the entire list of responses, falls within the 95% confidence interval calculated by each method for its particular range of responses. These were then used to give the observed readmission rates as detailed in the manuscript and results. **Table S3** gives an example of this for 30-day all-cause readmissions and the boundaries presented by RF.

Cohort Results

For the additional patients eliminated from the analysis because they died before being readmitted, we evaluated characteristics versus the cohort used for training. The analytic sample did not differ significantly from those who were excluded for various reasons (**Table S4**). Further, as the study excluded a large number of the original 1653 participants, we have listed the differences between the included cohort (n=1004) and excluded cohort (n=649) (**Table S5**). While many of the values are similar it might be interesting to further analyze their differences in future work. Finally, for the included patients, the percent missing for each variable is plotted in **Figure S1**. The complete missing information can be found in **Table S6.** In particular, the high rates of missing values for a large number of variables seems to be in large part as a result of incomplete questionnaires leading to summary scores that could not be calculated. While no individual patient within the cohort selected has a large rate of missing data, it does seem that the variables missing are consistent across all of the patients, which improves the likelihood that imputation is not causing a large effect on the outcome.

SUPPLEMENTAL REFERENCES

1. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 2010;33:1.
2. Zhang T. On the consistency of feature selection using greedy least squares regression. *J Mach Learn Res*. 2009;10:555-568.
3. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C and Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:1.
4. Liaw A and Wiener M. Classification and regression by randomForest. *R news*. 2002;2:18-22.

5. Culp M, Johnson K and Michailidis G. ada: An r package for stochastic boosting. *J Stat Softw*. 2006;17:9.

6. Chang C-C and Lin C-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2011;2:27.

7. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.

8. Hapfelmeier A and Ulm K. A new variable selection approach using random forests. *Comput Stat Data Anal*. 2013;60:50-69.

9. Genuer R, Poggi J-M and Tuleau-Malot C. Variable selection using random forests. *Pattern Recognition Letters*. 2010;31:2225-2236.

10. Verikas A, Gelzinis A and Bacauskiene M. Mining data with random forests: a survey and results of new tests. *Pattern Recognition*. 2011;44:330-349.

SUPPLEMENTAL FIGURE LEGEND

Figure S1. Variable missing rates for all variables with a rate of missing greater than 3% across the 1004 patient cohort.

SUPPLEMENTAL TABLES

Table S1. The 236 Features Used in the Tele-HF Analysis

Feature	Notes
Patient Information	
Age Over 90	Age greater than 90?
Age	Age of Patient
Age De-Identified	De-identified Age
Payor Type 1	Insurance: Commercial/PPO
Payor Type 3	Insurance: HMO
Payor Type 2	Insurance: Medicaid
Payor Type 5	Insurance: Medicare
Payor Type 6	Insurance: None/Self-Pay
Payor Type 7	Insurance: Other
Payor Type 8	Insurance: Unknown
Payor Type 4	Insurance: VA
Telemonitoring	Is patient assigned to the telemonitoring group?
HF HOSP	Is the number of times admitted due to heart failure an estimate?
Ethnicity	Is the patient Hispanic, non-Hispanic, or unknown?
Rx1_0	Medication: On ACE Inhibitor or ARB
Rx2_0	Medication: On ARA
Rx3_0	Medication: On Beta Blocker
Rx4_0	Medication: On Digoxin
Rx5_0	Medication: On Loop Diuretics?
RACE_4	Race: American Indian/Alaska Native
RACE_3	Race: Asian

RACE_2	Race: Black/African-American
RACE_5	Race: Native Hawaiian/Pacific Islander
RACE_6	Race: Other
RACE_7	Race: Unknown
RACE_1	Race: White/Caucasian
HOSP	Number of hospital visits in past year
Admitted	Number of times admitted to hospital in past year
HF_HOSP	Number of times admitted to hospital in past year for heart failure
PATIENT_SCALE	Patient has their own weight scale
PATID	Patient ID in Trial
SEX	Patient Sex

Hospitalization Information

Symptom: Chest	Cause of Hospitalization: Chest Pain
Symptom: Fatigue	Cause of Hospitalization: Fatigue
Symptom: Heart	Cause of Hospitalization: Irregular Heart Beat or Palpitations
Symptom: Other	Cause of Hospitalization: Other
Symptom: Breath	Cause of Hospitalization: Shortness of breath
Symptom: Swelling	Cause of Hospitalization: Swelling

Patient History

Connective Tissue	History: Connective Tissue Disease
Coronary Artery	History: Coronary Artery Disease
Diabetes	History: Diabetes
Diabetes: Organ	History: Diabetes with end organ damage
AIDS	History: Does patient have AIDS?
Hemiplegia	History: does patient have hemiplegia?

Tumor	History: Does patient have history of tumor?
Apnea	History: Does patient have sleep apnea?
Liver Disease Rating	History: if patient has liver disease, is it mild or moderate/severe?
Dialysis	History: On Dialysis
Leukemia	History: patient has a history of leukemia
Liver Disease	History: Patient has a history of liver disease
Lymphoma	History: Patient has a history of lymphoma
Metastatic Tumor	History: Patient has a history of metastatic tumors
Pacemaker	History: Patient has a permanent pacemaker
Chronic renal failure	History: Patient has chronic renal failure
Cardio Resync	History: Patient has had cardio resynchronization therapy
Cerebrovascular Disease	History: Patient has had cerebrovascular disease
Chronic Pulmonary	History: Patient has had chronic pulmonary disease
Hypercholesterolemia	History: Patient has history of hypercholesterolemia
Hypertension	History: Patient has history of hypertension
Drugs	History: Patient has history of illicit drug use
Ischemic Cardiomyopathy	History: Patient has history of ischemic cardiomyopathy
Peptic Ulcer	History: Patient has peptic ulcer disease
PVD	History: Patient has peripheral vascular disease
Aortic	History: Patient has severe aortic or mitral valve disease
Disease Management	History: Patient is in another disease management program
Oxygen	History: Patient uses Oxygen at home
Chronic Renal: Type	History: Patient with chronic renal failure has either mild or moderate/severe
Dementia	History: Patient has dementia

AICD	History: Prior AICD implantation
MI	History: Prior MI
TIA	History: Prior TIA

Laboratory Values and Physical Exams

Albumin	Laboratory Value: Albumin
Blood Urea Nitrogen	Laboratory Value: Blood Urea Nitrogen
Body Mass Index	Laboratory Value: Body Mass Index
BNP	Laboratory Value: Brain Natriuretic Peptide
CREATININE	Laboratory Value: Creatinine
DIASTOLIC_BP	Laboratory Value: Diastolic BP
GFR	Laboratory Value: Glomular Filtration Rate
HEIGHT	Laboratory Value: Height
HEMOCRIT	Laboratory Value: Hemocrit
HEMOGLOBIN	Laboratory Value: Hemoglobin
HIP	Laboratory Value: Hip circumference
JVD_MEASURE	Method by which JVD was calculated
JVD	Laboratory Value: Jugular Venous Distension
LVEF_METHOD	Laboratory Value: method by which LVEF was calculated
NT_PRO_BNP	Laboratory Value: N-terminal pro b-type Natriuretic Peptide
NYHA	Laboratory Value: NYHA Class
LOW_COG	Laboratory Value: Patient's Folstein score was less than or equal to 24
LVEF_PERCENT	Laboratory Value: Percentage of LVEF
PITTING_EDEMA	Laboratory Value: Pitting Edema (yes, no, or unsure)
POTASSIUM	Laboratory Value: Potassium

S3_PRESENCE	Laboratory Value: Presense of S3 (yes, no, unsure)
PULMONARY_RALES	Laboratory Value: Pulmonary Rales (Base, Above, or Clear)
RESP_RATE	Laboratory Value: Respiratory Rate
RESTING_HR	Laboratory Value: Resting Heart Rate
SYSTOLIC_BP	Laboratory Value: Systolic Blood Pressure
TEMP	Laboratory Value: Temperature
WAIST	Laboratory Value: Waist Circumference
LVEF40	Laboratory Value: Was patient's LVEF under 40%?
WEIGHT	Laboratory Value: Weight
Folstein	Folstein mini mental status exam score
Hemorrhagic	Hemorrhagic type: None, yes, or no
Surveys	
Contribute	Did someone other than the patient complete the baseline survey?
Help Complete	Did the patient have someone help complete surveys?
Work for Pay	Does patient currently work for pay?
Alone	Does patient live alone?
Doctor	Does the patient currently have a primary care provider?
Insurance	Does the patient have insurance?
Smoke	Does the patient smoke? And if so how often?
Financial	Financially, how comfortable is patient's household on patient's income?
Difficult	How difficult is it for the patient to receive health care?
Follow Up	How good is the patient's doctor at following up?
Smoke QT	How many cigarettes in the past 30 days?
Doc Days	How many days has patient been seeing current doctor?

Doc Mos	How many months has patient been seeing current doctor?
Dependent	How many people are dependent upon patient's income?
Doc Yrs	How many years has patient been seeing current doctor?
Medcosts	How often did patient avoid taking medication due to costs?
Smoke Age	How old was patient when she/he started smoking regularly?
Knowledge	How well does patient know own health?
Patient Scale: Source	If patient has a weight scale, did trial provide it?
Visits per week	If patient has been visited by home health care, how many visits per week?
Visits QT	If patient has been visited by home health care, how many visits?
Health	In general, how does the patient feel about her/his own health?
Visits	In past 3 months has patient been visited by home health care?
Avoided	In past year, has patient avoided getting health care because of cost?
Burden	In Past year, how much of a financial burden have medical costs been?
Studies	Is patient involved in other studies?
Religion	Is religion a source of strength for the patient?
INHOSPITAL	Was the patient's interview/screen conducted in a hospital?
EDUCATION	What is the highest level of education the patient has completed?
INCOME	What is the patient's total household income?
ENROLL_LANGUAGE	What language did the patient enroll in?
HOME	Where does the patient currently live?
ESSI	
ESSI: Close	ESSI: How often does patient have contact with someone close to

	them?
ESSI: Count On	ESSI: How often does patient have someone she/he can count on to listen?
ESSI: Chores	ESSI: How often does patient have someone to help with chores?
ESSI: Support	ESSI: How often does patient have someone who can provide emotional support?
ESSI: Advice	ESSI: How often does patient have someone who is able to provide good advice?
ESSI: Affection	ESSI: How often does patient have someone who shows them affection?
ESSI	Summary Score of the ESSI Survey
KCCQ	
KCCQ_WORK	KCCQ: How much does heart failure affect working or chores?
KCCQ_VISITING	KCCQ: How much does your heart failure affect visiting family?
KCCQ_HOBBIES	KCCQ: How much has heart failure affected your hobbies?
KCCQ_INTIMATE	KCCQ: How much has heart failure affected your intimate relationships?
KCCQ_LIFE	KCCQ: How satisfied with the rest of life would patient be with this level of heart failure?
KCCQ_CALL	KCCQ: how sure is patient in knowing who to call if heart failure gets worse?
KCCQ_UNDERSTAND	KCCQ: How well do you understand how to keep from getting worse?
KCCQ_SYMPTOMS	KCCQ: In past 2 weeks, have your symptoms changed?
KCCQ_LIMITED	KCCQ: In past 2 weeks, how much has heart failure limited your

	enjoyment of life?
KCCQ_FATIGUEBOTHER	KCCQ: In past 2 weeks, how often has fatigue bothered you?
KCCQ_FATIGUE	KCCQ: In past 2 weeks, how often has fatigue limited you?
KCCQ_YARDWORK	KCCQ: In past 2 weeks, how often has heart failure limited yard work?
KCCQ_DISCOURAGED	KCCQ: In past 2 weeks, how often has patient felt discouraged?
KCCQ_SHORTNESSBOTHER	KCCQ: In past 2 weeks, how often has shortness of breath bothered you?
KCCQ_SHORTNESS	KCCQ: In past 2 weeks, how often has shortness of breath limited you?
KCCQ_SLEEPING	KCCQ: In past 2 weeks, how often have you slept sitting up?
KCCQ_SWELLING	KCCQ: In past 2 weeks, how often have you woken up with swelling?
KCCQ_BATHING	KCCQ: In past two weeks, how limited by heart failure has bathing been?
KCCQ_DRESSING	KCCQ: In past two weeks, how limited by heart failure has dressing yourself been?
KCCQ_JOGGING	KCCQ: In past two weeks, how limited by heart failure has jogging been?
KCCQ_STAIRS	KCCQ: In past two weeks, how limited by heart failure has stair climbing been?
KCCQ_WALKING	KCCQ: In past two weeks, how limited by heart failure has walking one block been?
KCCQ_BOTHER	KCCQ: In past two weeks, how much has swelling bothered you?
KCCQ_PHYSICAL	KCCQ: Physical Limitations Summary Score

KCCQ_QUALITY	KCCQ: Quality of Life Summary Score
KCCQ_EFFICACY	KCCQ: Self-Efficacy Summary Score
KCCQ_SOCIAL	KCCQ: Social Limitations Summary Score
KCCQ_SUMMARY	KCCQ: Summary Score
KCCQ_BURDEN	KCCQ: Symptom Burden Summary Score
KCCQ_FREQUENCY	KCCQ: Symptom Frequency Summary Score
KCCQ_STABILITY	KCCQ: Symptom Stability Summary Score
KCCQ_SYMPSCORE	KCCQ: Symptom Summary Score
Morisky	
MORISKY_FORGOT	Morisky: Has patient forgotten to take medicine?
MORISKY_STOP	Morisky: Has patient stopped taking medication when feeling better?
MORISKY_WORSE	Morisky: Has patient stopped taking medication when feeling worse?
MORISKY_CARELESS	Morisky: Is Patient careless about taking medicine?
MISSMORISKY	Patient's Morisky Miss Summary Score
MORISKY	Patient's Morisky Summary Score
PHQ9	
PHQ9_INTEREST	PHQ9: In past 2 weeks, how often had litter interest or pleasure doing things?
PHQ9_TIRED	PHQ9: In past two weeks, how often feeling tired?
PHQ9_SLOW	PHQ9: In past two weeks, how often moving or speaking slower than usual?
PHQ9_SLEEPING	PHQ9: In past two weeks, how often trouble sleeping?
PHQ9_FAILURE	PHQ9: Over last two weeks, how often felt bad or a failure?

PHQ9_DEPRESSED	PHQ9: Over last two weeks, how often felt down or depressed?
PHQ9_BOD	PHQ9: Over last two weeks, how often felt would be better off dead?
PHQ9_APPETITE	PHQ9: Over last two weeks, how often troubled by poor appetite?
PHQ9_CONCENTRATING	PHQ9: Over last two weeks, how often troubled concentrating?
PHQ9	PHQ9: Summary Score
PSS	
PSS_DIFFICULTY	PSS: In past month, difficulties were piling up
PSS_YOURWAY	PSS: in past month, felt things were going patient's way
PSS_CONTROL	PSS: In past month, felt unable to control important things?
PSS_CONFIDENT	PSS: In past month, how confident can handle personal problems?
PSS	PSS: Summary Score
REALM	
REALM_NONE	REALM: could not read any terms
REALM_ALLERGIC	REALM: could read Allergic?
REALM_ANEMIA	REALM: could read Anemia?
REALM_COLITIS	REALM: could read Colitis?
REALM_CONSTIPATION	REALM: could read Constipation?
REALM_DIRECTED	REALM: could read Directed?
REALM_FAT	REALM: could read Fat?
REALM_FATIGUE	REALM: could read Fatigue?
REALM_FLU	REALM: could read Flu?
REALM_JAUNDICE	REALM: could read Jaundice?

REALM_OSTEOPOROSIS	REALM: could read Osteoporosis?
REALM_PILL	REALM: could read Pill?
REALM_SCORE	REALM: Summary Score
SCHFI	
SCHFI_BETTER	SCHFI: How sure are you that you can do something to make symptoms better?
SCHFI_SERIOUS	SCHFI: How sure are you that you can judge how serious symptoms are?
SCHFI_JUDGE	SCHFI: How sure are you that you can judge what makes symptoms better?
SCHFI_HEALTH	SCHFI: How sure are you that you can notice changes in health?
SCHFI	SCHFI: Summary Score
WARE	
WARE_NEED	WARE: Can get medical care when patient needs it?
WARE_FINANCIAL	WARE: Can get medical care without financial setback?
WARE_EXPLAIN	WARE: Doctor explains things well?
WARE_TIME	WARE: Doctor spends plenty of time with patient?
WARE_FRIENDLY	WARE: Doctor treats patient in friendly manner?
WARE_IGNORE	WARE: Doctors ignores what patient tells them?
WARE_OFFICE	WARE: Doctor's office has everything patient needs?
WARE_DOUBTS	WARE: Doubts about how your doctor is treating you?
WARE_SPECIALIST	WARE: Easy access to medical specialist when needed?
WARE_CAREFUL	WARE: How careful is your doctor to check everything?
WARE_DISSATISFIED	WARE: How dissatisfied with things in your medical care are you?

WARE_APPOINTMENT	WARE: how hard is it to get an appointment?
WARE_CARE	WARE: How perfect do you feel your medical care is?
WARE_1	WARE: intermediate score
WARE_2	WARE: intermediate score
WARE_3	WARE: intermediate score
WARE_4	WARE: intermediate score
WARE_5	WARE: intermediate score
WARE_6	WARE: intermediate score
WARE_7	WARE: intermediate score
WARE_EMERGENCY	WARE: Medical care takes too long for emergency treatment?
WARE_HURRY	WARE: Providers hurry too much when treating patient?
WARE	WARE: Summary Score
WARE_AFFORD	WARE: With care being received now, how well do you feel you can afford care?
WARE_DOCTOR	WARE: Your doctor acts too business like?

Table S2. Variables Selected in the Forward, Stepwise Selection Method for LR

Feature	Number of Times Selected
ALBUMIN_	1
METASTATIC_TUMOR_	5
CORONARY_ARTERY_	3
KCCQ_BOTHER_	3
ALONE_	1
WARE_1_	1
PHQ9_INTEREST_	3
CHRONIC_RENAL_FAILURE_	1
DEPENDENT_	1
WARE_EXPLAIN_	1
BLOOD_UREA_NITROGEN_	6
KCCQ_EFFICACY_	6
KCCQ_DISCOURAGED_	1
PATIENT_SCALE_SOURCE_	1
WARE_	3
DIABETES_	1
DOCYRS_	2
RACE_6_	1
PHQ9_	1
CONNECTIVE_TISSUE_	2
KCCQ_QUALITY_	1
KCCQ_STAIRS_	4
KCCQ_UNDERSTAND_	3

ENROLL_LANGUAGE_	2
KCCQ_FREQUENCY_	3
WARE_EMERGENCY_	1
CREATININE_	2
JVD_MEASURE_	1
BURDEN_	17
RELIGION_	9
MORISKY_FORGOT_	4
KCCQ_JOGGING_	12
KCCQ_BURDEN_	5
ESSI_CHORES_	6
REALM_DIRECTED_	1
MEDCOSTS_	2
KCCQ_VISITING_	2
CHRONIC_PULMONARY_	5
ADMITTED_	28
SCHFI_	3
REALM_OSTEOPOROSIS_	1
KCCQ_SHORTNESSBOTHER_	2
KCCQ_STABILITY_	1
KCCQ_WORK_	4
FINANCIAL_	1
DISEASE_MANAGEMENT_	1
WARE_SPECIALIST_	2
PAYOR_8_	2

KCCQ_WALKING_	1
GFR_	2
KCCQ_SWELLING_	24
RACE_3_	6
Rx1_0_	1
INSURANCE_	2
PSS_DIFFICULTY_	1
PAYOR_6_	2
KCCQ_SUMMARY_	5
OXYGEN_	1
Rx2_0_	1
DOCMOS_	1
HEALTH_	2
EDUCATION_	1
PEPTIC_ULCER_	1
LIVER_DISEASE_RATING_	1
KCCQ_HOBBIES_	1
KCCQ_YARDWORK_	5
HF_HOSP_	13
KCCQ_PHYSICAL_	4
KCCQ_SYMPSCORE_	1
SCHFI_HEALTH_	1
TUMOR_	2
VISITEDQT_	2
KCCQ_FATIGUE_	2

HELP_COMPLETE_	3
POTASSIUM_	1
MORISKY_WORSE_	1
MORISKY_STOP_	4
KCCQ_SYMPTOMS_	4
PRIOR_AICD_	7
HF_HOSP_EST_	24

Table S3. Deciles and Confidence Intervals for 30-day All-Cause Readmission Responses from RF

Deciles	Overall Boundary	Mean (95% CI)
1	0.392	0.393 (0.390-0.396)
2	0.418	0.420 (0.417-0.423)
3	0.438	0.439 (0.436-0.442)
4	0.456	0.456 (0.453-0.459)
5	0.472	0.472 (0.469-0.475)
6	0.488	0.487 (0.484-0.490)
7	0.504	0.503 (0.500-0.506)
8	0.524	0.522 (0.519-0.525)
9	0.550	0.548 (0.545-0.551)
10	1.00	1.00 (1.00-1.00)

Table S4. Excluded Patients with Death but No Readmission from 180-Day Analysis Set

Characteristic	180-Day	
	Excluded **N (%)**	**Included** **N (%)**
N	27 (100.0)	977 (100.0)
Median age (SD)	66.5 (12.2)	62 (15.7)
Females	12 (44.4)	403 (41.2)
Race		
White	16 (59.3)	491 (50.3)
African American	8 (29.6)	385 (39.4)
Other	3 (11.1)	101 (10.3)
New York Heart Association		
Class I	2 (7.41)	54 (5.5)
Class II	15 (55.6)	500 (51.2)
Class III	8 (29.6)	347 (35.5)
Class IV	1 (3.70)	57 (5.8)
Missing	1 (3.70)	19 (2.0)
Medical History		
LVEF† % < 40	19 (70.4)	668 (68.4)
Hypertension	19 (70.4)	752 (77.0)
Diabetes	11 (40.7)	439 (44.9)
Myocardial Infarction	7 (25.9)	250 (25.6)
Stroke	4 (14.8)	92 (9.4)
Ischemic	7 (25.9)	228 (23.3)

Cardiomyopathy		
Clinical Values (Mean/SD)		
Albumin	3.5 (0.37)	3.31 (0.53)
Blood Urea Nitrogen	36.4 (26.5)	26.3 (16.8)
Creatinine	1.70 (1.21)	1.45 (0.72)
Hemoglobin	11.4 (1.75)	12.4 (1.94)
Glomerular Filtration Rate	49.3 (22.7)	58.8 (27.4)
Potassium	3.93 (0.59)	4.08 (0.57)

*All values in tables are mean (standard deviation) unless noted.

†LVEF: Left Ventricular Ejection Fraction.

Table S5. Excluded Patient Characteristics

Characteristic	Included N (%)	Excluded N (%)
N	1004	649
Readmitted over 180 days (Rate)	478 (47.6)	321 (49.5)
Median age (SD)	62 (15.7)	59 (16.4)
Females	415 (41.3)	280 (43.1)
Race		
White	507 (50.5)	309 (47.6)
African American	393 (39.1)	251 (38.7)
Other	104 (10.4)	89 (13.7)
New York Heart Association		
Class I	56 (5.6)	48 (7.4)
Class II	515 (51.3)	309 (47.6)
Class III	355 (35.4)	243 (37.4)
Class IV	58 (5.8)	41 (6.3)
Missing	20 (1.9)	8 (1.2)

Medical History		
LVEF† % <40	687 (68.4)	448 (69.0)
Hypertension	771 (76.8)	477 (73.5)
Diabetes	450 (44.8)	197 (30.3)
Myocardial Infarction	257 (25.6)	143 (22.0)
Stroke	96 (9.6)	53 (8.2)
Ischemic Cardiomyopathy	235 (23.4)	141 (21.7)
Clinical Values (Mean/SD)		
Albumin	3.32 (0.53)	2.01 (1.65)
Blood Urea Nitrogen	25.2 (17.8)	27.1 (18.4)
Creatinine	1.40 (0.77)	1.44 (0.76)
Hemoglobin	12.3 (1.94)	11.5 (3.88)
Glomerular Filtration Rate	58.5 (27.4)	52.8 (31.1)
Potassium	4.08	3.93

	(0.57)	(0.92)

*All values are mean (standard deviation) unless noted.

†LVEF, Left Ventricular Ejection Fraction

Table S6. Percentage of Missing Per Variable from the 236 Variables in the Cohort of 1004 Patients

Variable	% Missing
AGE_DI_MISSING	0.00
AGEOVR90_MISSING	0.00
AORTICDISEASE_MISSING	0.00
DIABETES_MISSING	0.00
ENROLL_LANGUAGE_MISSING	0.00
ESSI_MISSING	0.00
ETHNICITY_MISSING	0.00
FOLSTEIN_SCORE_MISSING	0.00
GROUP_MISSING	0.00
INHOSPITAL_MISSING	0.00
LOW_COG_MISSING	0.00
LVEF40_MISSING	0.00
MORISKY_CARELESS_MISSING	0.00
MORISKY_FORGOT_MISSING	0.00
MORISKY_STOP_MISSING	0.00
MORISKY_WORSE_MISSING	0.00
PAYOR_1_MISSING	0.00
PAYOR_2_MISSING	0.00
PAYOR_3_MISSING	0.00
PAYOR_4_MISSING	0.00
PAYOR_5_MISSING	0.00
PAYOR_6_MISSING	0.00
PAYOR_7_MISSING	0.00

PAYOR_8_MISSING	0.00
PHQ9_MISSING	0.00
RACE_1_MISSING	0.00
RACE_2_MISSING	0.00
RACE_3_MISSING	0.00
RACE_4_MISSING	0.00
RACE_5_MISSING	0.00
RACE_6_MISSING	0.00
RACE_7_MISSING	0.00
Rx1_0_MISSING	0.00
Rx2_0_MISSING	0.00
Rx3_0_MISSING	0.00
Rx4_0_MISSING	0.00
Rx5_0_MISSING	0.00
SEX_MISSING	0.00
SYMP_BREATH_MISSING	0.00
SYMP_CHEST_MISSING	0.00
SYMP_FATIGUE_MISSING	0.00
SYMP_HEART_MISSING	0.00
SYMP_OTHER_MISSING	0.00
SYMP_SWELL_MISSING	0.00
WARE_1_MISSING	0.00
WARE_2_MISSING	0.00
WARE_3_MISSING	0.00
WARE_4_MISSING	0.00

WARE_6_MISSING	0.00
WARE_MISSING	0.00
KCCQ_QUALITY_MISSING	0.10
BMI_MISSING	0.20
ESSI_AFFECTION_MISSING	0.20
KCCQ_EFFICACY_MISSING	0.20
KCCQ_SYMPSCORE_MISSING	0.20
MISSMORISKY_MISSING	0.20
ESSI_CHORES_MISSING	0.30
ESSI_COUNTON_MISSING	0.30
ESSI_ADVICE_MISSING	0.40
KCCQ_HOBBIES_MISSING	0.40
WARE_7_MISSING	0.50
KCCQ_STAIRS_MISSING	0.70
PHQ9_DEPRESSED_MISSING	0.70
WARE_EXPLAIN_MISSING	0.70
MORISKY_MISSING	0.80
PHQ9_CONCENTRATING_MISSING	0.80
KCCQ_WORK_MISSING	0.90
WARE_5_MISSING	0.90
WARE_OFFICE_MISSING	0.90
EDUCATION_MISSING	1.00
ESSI_CLOSE_MISSING	1.00
KCCQ_INTIMATE_MISSING	1.00
KCCQ_SOCIAL_MISSING	1.00

KCCQ_YARDWORK_MISSING	1.00
PHQ9_SLEEPING_MISSING	1.00
AGE_MISSING	1.10
DIFFICULT_MISSING	1.10
INSURANCE_MISSING	1.10
KCCQ_BATHING_MISSING	1.10
KCCQ_BURDEN_MISSING	1.10
KCCQ_SYMPTOMS_MISSING	1.10
PHQ9_SLOW_MISSING	1.10
WARE_CARE_MISSING	1.10
ESSI_EMOTIONAL_MISSING	1.20
KCCQ_DRESSING_MISSING	1.20
KCCQ_VISITING_MISSING	1.20
PHQ9_FAILURE_MISSING	1.20
PHQ9_INTEREST_MISSING	1.20
PHQ9_TIRED_MISSING	1.20
WARE_DOCTOR_MISSING	1.20
WARE_FRIENDLY_MISSING	1.20
MEDCOSTS_MISSING	1.29
WARE_AFFORD_MISSING	1.29
KCCQ_JOGGING_MISSING	1.39
PHQ9_APPETITE_MISSING	1.39
SMOKE_MISSING	1.39
WARE_IGNORE_MISSING	1.39
HEALTH_MISSING	1.49

KCCQ_SHORTNESS_MISSING	1.49
KCCQ_WALKING_MISSING	1.59
WARE_HURRY_MISSING	1.59
KCCQ_DISCOURAGED_MISSING	1.69
KCCQ_FATIGUE_MISSING	1.69
KCCQ_LIMITED_MISSING	1.69
KCCQ_SLEEPING_MISSING	1.69
PHQ9_BOD_MISSING	1.69
WARE_FINANCIAL_MISSING	1.69
WARE_SPECIALIST_MISSING	1.69
WORKPAY_MISSING	1.69
BURDEN_MISSING	1.79
KCCQ_UNDERSTAND_MISSING	1.79
WARE_NEED_MISSING	1.79
HOME_MISSING	1.89
WARE_TIME_MISSING	1.89
AIDS_MISSING	1.99
APNEA_MISSING	1.99
CARDIO_RESYNC_MISSING	1.99
CEREBROVASCULAR_DISEASE_MISSING	1.99
CHRONIC_PULMONARY_MISSING	1.99
CHRONIC_RENAL_FAILURE_MISSING	1.99
CONNECTIVE_TISSUE_MISSING	1.99
CORONARY_ARTERY_MISSING	1.99
DIABETES_ORGAN_MISSING	1.99

DIALYSIS_MISSING	1.99
DIASTOLIC_BP_MISSING	1.99
DIMENTIA_MISSING	1.99
DISEASE_MANAGEMENT_MISSING	1.99
DRUG_USE_MISSING	1.99
HEIGHT_MISSING	1.99
HELP_COMPLETE_MISSING	1.99
HEMIPLEGIA_MISSING	1.99
HIP_MISSING	1.99
HYPERCHOLESTEROLEMIA_MISSING	1.99
HYPERTENSION_MISSING	1.99
ISCHEMIC_CARDIOMYOPATHY_MISSING	1.99
JVD_MISSING	1.99
LEUKEMIA_MISSING	1.99
LIVER_DISEASE_MISSING	1.99
LVEF_METHOD_MISSING	1.99
LYMPHOMA_MISSING	1.99
METASTATIC_TUMOR_MISSING	1.99
NYHA_MISSING	1.99
OXYGEN_MISSING	1.99
PACEMAKER_MISSING	1.99
PATIENT_SCALE_MISSING	1.99
PEPTIC_ULCER_MISSING	1.99
PERIPH_VASCULAR_DISEASE_MISSING	1.99
PITTING_EDEMA_MISSING	1.99

PRIOR_AICD_MISSING	1.99
PRIOR_MI_MISSING	1.99
PRIOR_TIA_MISSING	1.99
PULMONARY_RALES_MISSING	1.99
REALM_ALLERGIC_MISSING	1.99
REALM_ANEMIA_MISSING	1.99
REALM_COLITIS_MISSING	1.99
REALM_CONSTIPATION_MISSING	1.99
REALM_DIRECTED_MISSING	1.99
REALM_FAT_MISSING	1.99
REALM_FATIGUE_MISSING	1.99
REALM_FLU_MISSING	1.99
REALM_JAUNDICE_MISSING	1.99
REALM_NONE_MISSING	1.99
REALM_OSTEOPOROSIS_MISSING	1.99
REALM_PILL_MISSING	1.99
REALM_SCORE_MISSING	1.99
RESP_RATE_MISSING	1.99
RESTING_HR_MISSING	1.99
S3_PRESENCE_MISSING	1.99
SYSTOLIC_BP_MISSING	1.99
TEMP_MISSING	1.99
TUMOR_MISSING	1.99
WAIST_MISSING	1.99
WARE_CAREFUL_MISSING	1.99

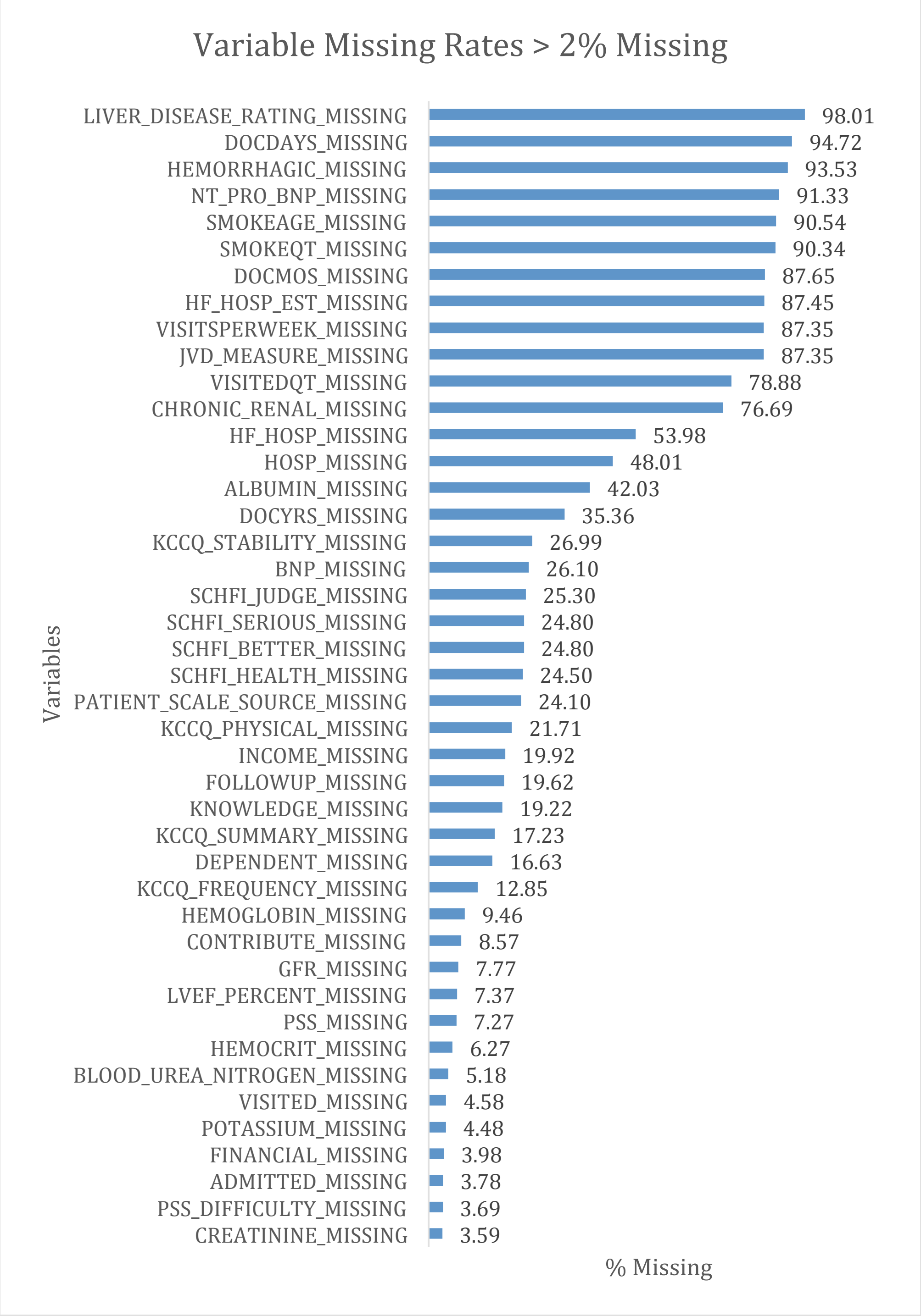
WEIGHT_MISSING	1.99
AVOIDED_MISSING	2.09
KCCQ_BOTHER_MISSING	2.09
KCCQ_SHORTNESSBOTHER_MISSING	2.09
KCCQ_SWELLING_MISSING	2.09
WARE_DISSATISFIED_MISSING	2.09
DOCTOR_MISSING	2.19
PSS_CONTROL_MISSING	2.19
PSS_CONFIDENT_MISSING	2.39
STUDIES_MISSING	2.39
WARE_EMERGENCY_MISSING	2.39
WARE_APPOINTMENT_MISSING	2.49
KCCQ_FATIGUEBOTHER_MISSING	2.59
PSS_YOURWAY_MISSING	2.59
SCHFI_MISSING	2.59
KCCQ_CALL_MISSING	2.69
KCCQ_LIFE_MISSING	2.69
RELIGION_MISSING	2.69
WARE_DOUBTS_MISSING	2.69
ALONE_MISSING	2.79
CREATININE_MISSING	3.59
PSS_DIFFICULTY_MISSING	3.69
ADMITTED_MISSING	3.78
FINANCIAL_MISSING	3.98
POTASSIUM_MISSING	4.48

VISITED_MISSING	4.58
BLOOD_UREA_NITROGEN_MISSING	5.18
HEMOCRIT_MISSING	6.27
PSS_MISSING	7.27
LVEF_PERCENT_MISSING	7.37
GFR_MISSING	7.77
CONTRIBUTE_MISSING	8.57
HEMOGLOBIN_MISSING	9.46
KCCQ_FREQUENCY_MISSING	12.85
DEPENDENT_MISSING	16.63
KCCQ_SUMMARY_MISSING	17.23
KNOWLEDGE_MISSING	19.22
FOLLOWUP_MISSING	19.62
INCOME_MISSING	19.92
KCCQ_PHYSICAL_MISSING	21.71
PATIENT_SCALE_SOURCE_MISSING	24.10
SCHFI_HEALTH_MISSING	24.50
SCHFI_BETTER_MISSING	24.80
SCHFI_SERIOUS_MISSING	24.80
SCHFI_JUDGE_MISSING	25.30
BNP_MISSING	26.10
KCCQ_STABILITY_MISSING	26.99
DOCYRS_MISSING	35.36
ALBUMIN_MISSING	42.03
HOSP_MISSING	48.01

HF_HOSP_MISSING	53.98
CHRONIC_RENAL_MISSING	76.69
VISITEDQT_MISSING	78.88
JVD_MEASURE_MISSING	87.35
VISITSPERWEEK_MISSING	87.35
HF_HOSP_EST_MISSING	87.45
DOCMOS_MISSING	87.65
SMOKEQT_MISSING	90.34
SMOKEAGE_MISSING	90.54
NT_PRO_BNP_MISSING	91.33
HEMORRHAGIC_MISSING	93.53
DOCDAYS_MISSING	94.72
LIVER_DISEASE_RATING_MISSING	98.01

SUPPLEMENTAL FIGURE

Figure S1


Variable Missing Rates > 2% Missing
Variables
% Missing
LIVER_DISEASE_RATING_MISSING 98.01
DOCDAYS_MISSING 94.72
HEMORRHAGIC_MISSING 93.53
NT_PRO_BNP_MISSING 91.33
SMOKEAGE_MISSING 90.54
SMOKEQT_MISSING 90.34
DOCMOS_MISSING 87.65
HF_HOSP_EST_MISSING 87.45
VISITSPERWEEK_MISSING 87.35
JVD_MEASURE_MISSING 87.35
VISITEDQT_MISSING 78.88
CHRONIC_RENAL_MISSING 76.69
HF_HOSP_MISSING 53.98
HOSP_MISSING 48.01
ALBUMIN_MISSING 42.03
DOCYRS_MISSING 35.36
KCCQ_STABILITY_MISSING 26.99
BNP_MISSING 26.10
SCHFI_JUDGE_MISSING 25.30
SCHFI_SERIOUS_MISSING 24.80
SCHFI_BETTER_MISSING 24.80
SCHFI_HEALTH_MISSING 24.50
PATIENT_SCALE_SOURCE_MISSING 24.10
KCCQ_PHYSICAL_MISSING 21.71
INCOME_MISSING 19.92
FOLLOWUP_MISSING 19.62
KNOWLEDGE_MISSING 19.22
KCCQ_SUMMARY_MISSING 17.23
DEPENDENT_MISSING 16.63
KCCQ_FREQUENCY_MISSING 12.85
HEMOGLOBIN_MISSING 9.46
CONTRIBUTE_MISSING 8.57
GFR_MISSING 7.77
LVEF_PERCENT_MISSING 7.37
PSS_MISSING 7.27
HEMOCRIT_MISSING 6.27
BLOOD_UREA_NITROGEN_MISSING 5.18
VISITED_MISSING 4.58
POTASSIUM_MISSING 4.48
FINANCIAL_MISSING 3.98
ADMITTED_MISSING 3.78
PSS_DIFFICULTY_MISSING 3.69
CREATININE_MISSING 3.59