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