Web Only Supplement

eTable 1. Comorbidity Attributes.

Comorbidity Attributes	ICD-9 Codes	CPT Codes	Additional Descriptor
Hypertension[a]	401.x, 402.x, 403.x, 404.x, 405.x, 437.2		
Coronary Artery Disease[a]	410.xx, 411.xx, 412.xx, 414.xx	33410-22141, 33510-33536, 33570, 33575, 35600, S2204-2209, 92973, 92980-92984, 92995-92996, G0290, G0291, S2220, 92975-92977	
Myocardial Infarction	410.xx, 411.0, 412.xx		
Heart Failure[a,b]	398.91, 402.x1, 404.x1, 404.x3, 425.2x-452.9x, 428.0x-428.2x, 428.4x-428.9x		
Atrial Fibrillation[a]	427.31		
Non-Atrial Fibrillation Arrhythmias[a]	426.0, 426.1x, 427.0x-427.2x, 427.32, 427.4-427.9, V45.01, V53.3	33200-33249, 93536, G0297, G0298, G0299, G0300	
Diastolic Dysfunction[a]	428.3x, 428.4x, 429.9AL, 429.9BF, 429.9AR		
Valvular Disease[a]	093.2x, 391.1, 394.xx, 395.xx, 396.xx, 397.xx, 421.xx, 424.xx, 746.3-746.6, V42.2, V43.3	33400-33496, 92986-92990	
Pulmonary Hypertension[a]	416.0, 416.8, 416.9		
Peripheral Arterial Disease[a]	440.xx, 441.xx, 443.2x, 443.81, 443.9, 447.1, 557.1, 557.9	34520, 35501-35599, 35616-35623, 35637-35641, 35646-35661, 35665-35671	
Venous Thromboembolic Disease[a]	415.1x, 453.4x, V12.51		
Chronic Obstructive Pulmonary Disease[a,b]	491.2x, 492.xx		
Asthma[a,b]	493.xx		
Other Chronic Lung Disease[a]	494.xx-495.xx		
Obstructive Sleep Apnea[a]	327.2x, V46.8	94660, 95810	
Liver Disease – Severe	070.2x, 070.4x, 070.6x, 070.71, 571.1, 572.2-572.4, 456.0-456.1x	43243, 43244, 43400	
Liver Disease – Mild	070.3x, 070.5x, 070.70, 070.9, 571.0, 571.2-571.6, 571.8-571.9		Absence of Severe Liver Disease.
Liver Disease - Any[a]	070.2x-070.7x, 456.0-456.1x, 070.9, 571.0-571.9	43243, 43244, 43400	Mild or Severe Liver Disease.

eTable 1 (continued). Comorbidity Attributes.

Comorbidity Attributes	ICD-9 Codes	CPT Codes	Additional Descriptor
Pancreatitis[a]	577.0, 577.1		
Inflammatory Bowel Disease[a]	555.x-556.x		
Hemodialysis	996.1, 996.56, V45.1x, V56.x	36147,36148, 36800-36821, 36825-36830, 36832-36835, 37190, G0365, 90918-90940, 90998, G0257, G0308–G0319, 90945-90947, 90989-90997, 90999	
Chronic Kidney Disease	403.x, 404.x, 582.x, 583.x, 585.x-586.x, 588.x		
Any Chronic Kidney Disease[a]	403.x, 404.x, 582.x, 583.x, 585.x-586.x, 588.x, 996.1, 996.56, V45.1x, V56.x	36147,36148, 36800-36821, 36825-36830, 36832-36835, 37190, G0365, 90918-90940, 90998, G0257, G0308–G0319, 90945-90947, 90989-90997, 90999	Includes hemodialysis and/or chronic kidney disease
Diabetes Mellitus – Complicated	250.40-250.73, 250.90-250.93		
Diabetes Mellitus – Uncomplicated	250.0x-250.3x, 250.8x		
Diabetes Mellitus – Any[a]	250.x		Include Complicated and Uncomplicated Diabtese Mellitus.
Ischemic Stroke	282.61, 346.6x, 433.x1, 434.x1, 438.x, 997.02		
Transient Ischemic Attack	362.34, 435.9x, V12.54		
Intracranial Hemorrhage[a]	430.x-432.x		
Cerebrovascular Disease[a]	282.61, 346.6x, 362.34, 433.x-439.x, 997.02, V12.54	35301, 35390, 35701, 35901, 60600, 60605, G8240	Includes Ischemic stroke, and/or Transient Ischemic Attack.
Any Cardiovascular Disease	282.61, 346.6x, 362.34, 410.xx, 411.xx, 412.xx, 414.xx, 433.x-439.x, 440.xx, 441.xx, 443.2x, 443.81, 443.9, 447.1, 557.1, 557.9, 997.02, V12.54	33536, 33570, 33575, 34520, 35301, 35390, 35501-35599, 35600, 35616-35623, 35637-35641, 35646-35661, 35665-35671, 35701, 35901, 60600, 60605, 92973, 92980-92984, 92995-92996, 92975-92977, G0290, G0291, G8240, S2220, S2204-2209	Includes Coronary Artery Disease, Peripheral Arterial Disease, and/or Cerebrovascular Disease.
Hypothyroidism[a]	243.x-244.x		

eTable 1 (continued). Comorbidity Attributes.

Comorbidity Attributes	ICD-9 Codes	CPT Codes	Additional Descriptor
Hyperthyroidism[a]	242.x		
Any Thyroid Disorder	242.x-244.x		
Dementia[a]	290.0x-290.4x, 331.0-331.7x, 331.82, 331.9x, 046.1, 046.3		
Paralysis[a]	342.x-344.x, 438.2x-438.5x		
Myasthenia Gravis[a]	358.0x, 358.1x		
Parkinson's Disease[a]	332.x		
Multiple Sclerosis[a]	340.x		
Epilepsy[a]	345.xx		
Depression[a]	296.2x, 296.3x, 300.4, 311		
Bipolar Disorder[a]	296.0x-296.1x, 296.4x-296.8x		
Schizophrenia[a]	295.0x-295.9x		
Any mental illness (non-dementia)	291.xx-299.xx		
Any mental illness (includes dementia)	290.xx-299.xx		
Lupus	695.4, 710.0		
Scleroderma	701.0, 710.1		
Rheumatoid Arthritis	714.0x-714.4x, 714.81		
Polymyalgia Rheumatica	725.0		
Any collagen Vascular Disease[a]	695.4, 701.0, 710.x, 714.x, 720.x, 725.0		
HIV/AIDS[a]	V08, 042.x		
Non-cornea organ transplant[a]	238.77, 996.8x, V42.0-V42.4, V42.6-V42.9, V58.44, E878.0	38205-38215, 38231-38242, 86915, G0627, S2140, 50360-50365, 50380, S2065, 0014T, 32851-32856, 33930-33935, 33944-33945, 44135-44136, 47135-47147, 48160, 48551, 48554, 65780-65782, G0341- G0343, S2052-S2055, S2060, S2102, S2103, S2152	
Lung Cancer [b]	162.2x-162.9x		
Breast Cancer [b]	174.xx		
Prostate Cancer [b]	185.xx		
Cervical Cancer [b]	180.xx		
Colorectal Cancer [b]	153.xx, 154.xx		
Any Solid organ tumor[a,b]	140.xx-195.xx		
Leukemia [b]	204.xx-208.xx		
Lymphoma[b]	200.xx-202.xx		
Multiple Myeloma [b]	203.0x		
Any heme or lymphatic malignancy[a,b]	200.xx-208.xx		

eTable 1 (continued). Comorbidity Attributes.

Comorbidity Attributes	ICD-9 Codes	CPT Codes	Additional Descriptor
Metastatic Solid Tumor[b]	196.xx-199.xx		
Any malignancy[b]	140.xx-209.xx		
Abnormal mammogram	793.80, 793.81, 793.89		
Abnormal Pap test	795.00-795.06, 795.09		
Abnormal Colonoscopy	211.3, V12.72, V45.89	45385	
Coagulopathy[a]	286.0-286.9, 287.1, 287.3-287.5		
Anemia[a]	280.0-281.9, 285.0-285.		
Alcohol Abuse[a]	291.0x-291.9x, 303.9x, 305.0x, 648.4x		
Tobacco Use[a]	305.1, 649.0		Or patient identified as current smoker in social history.
Illicit Drug Use[a]	292.1x-292.9x, 304.xx, 305.2x-305.9x		Or patient identified as illicit drug user in social history.
Obesity[a]	278.00-278.01		
Weight Loss, Malnutrition, Anorexia[a]	263.0x-263.9x, 783.2x, 799.4		
Incontinence[a]	625.6, 788.3x, 787.6, 788.91		
Delirium[a]	290.11, 290.3, 290.41, 293.0, 293.1, 780.09		
Osteoporosis[a]	733.0x		
Osteoarthritis[a]	715.0x		
Any joint replacement		27420-27424, 27427-27429, 27437-27447, 27486-27487, 27125-27138, S2118, 0090T, 0091T, 0092T, 0096R-0098T, 21240-21243	
Gout[a]	274.0x		
Cataracts[a]	366.xx, 988.82		
Glaucoma[a]	365.xx		
Hearing Loss[a]	388.0x, 389.xx		
Oxygen Use	V46.2		
Fall[a,b]	E888.x		
Trauma	959.xx		
Any chemotherapy administration[b]	V58.1x		
Hospital Follow-up[b]	V67.59		

x=any number, a – Included in Comorbidity Count, b – Time Sensitive Encounter Diagnoses

eTable 2. Medication Attributes.

Count Medication Attributes	VA Class
Anti-hypertensives	AU200, CV100, CV200, CV400 (counts as two – combination pills), CV490, CV500, CV701, CV702, CV703, CV704, CV709, CV800, CV805
Steroids	HS051
Hypoglycemics	HS501 and HS502
Anticoagulants/Antiplatelets	BL110 and BL117
Antibiotics	AM550, AM600, AM650, AM700, AM900

eTable 3. Laboratory Attributes.

- Hemoglobin
- Albumin
- Potassium
- Sodium
- Bicarbonate
- Glucose
- Calcium
- Phosphorus
- Aspartate Aminotransferase
- Alanine Aminotransferase
- Alkaline Phosphatase
- Total Bilirubin
- International Normalized Ratio
- Blood Urea Nitrogen
- Creatinine
- Glomerular Filtration Rate
- Total Cholesterol
- High Denisty Lipoprotein
- Low Density Lipoprotein
- Triglycerides
- Brain natriuretic peptide
- Troponin
- Hemoglobin A1c
- Uric Acid

Mean, median, standard deviation, high, and low values for the year prior to the index visit were extracted.

eMethods

Feature Selection

The goal of feature selection is to reduce the number of attributes to be used in the model, while trying to retain the predictive power of the original set of attributes in the preprocessed data. We use the *Correlation Feature Selection (CFS)* [1] to identify a subset of attributes which were highly correlated with the outcome variable while having low inter-correlation amongst themselves. The CFS technique was used in conjunction with a greedy stepwise search to find the subset *S* with the best average merit, which is given by:

$$Merit_S = \frac{n\overline{r_{fo}}}{\sqrt{n + n(n-1)\overline{r_{ff}}}}$$

where *n* is the number of features in *S*, $\overline{r_{fc}}$ is the average value of feature-outcome correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations.

The relative predictive power of the final 24 features used in the model was assessed using the information gain metric, which evaluates the worth of an attribute by measuring the information gain with respect to the class:

$$InfoGain(Class, Attribute) = H(Class) - H(Class \mid Attribute)$$

where *H( )* denotes the information entropy.

Although we used cross validation for evaluating predictive models (discussed later), we used the entire dataset for feature selection, which can potentially bias the results and should be avoided in general (cross-validation should be used for feature selection as well). However, our obhservations below suggest that the bias is minimal in this case. The reason to use the entire dataset for feature selection process is our multi-step strategy, which included a manual screening to eliminate redundant features, etc. (as described in the manuscript). The goal at each step of feature selection was to get a single subset of features for the next step so as to eventually get a single subset for use in the final model. Using cross-validation for CFS would give slightly different subsets for each fold, which would complicate the manual screening step, and each resulting subset would again give different subsets after the second round of CFS. To simplify the process, we used the entire data at each step and got a single subset of features for the final model. This may however introduce some optimistic bias in model performance. To get a rough idea of the possible bias introduced in our model, we made two independent observations: i) We performed feature selection (CFS) using 10-fold cross-validation on the original dataset with 979 features, and found that each attribute in our final subset of 24 attributes appears in at least 7 out of the 10 subsets obtained from cross-validation (except 'sex' because it was manually added later); ii) We built the ensemble model on the original dataset using all 979 features with the goal of eliminating any possible bias due to feature selection. This model resulted in practically the same cross-validation c-statistic as that using the final 24 features only (0.854 without feature selection vs. 0.858 with feature selection, with a comparison p-value of 0.2085). Both these observations lead us to believe that our chosen strategy for feature selection introduces minimal bias in the final results.

Predictive Modeling

The outcome prediction database consisted of 7463 instances and 25 attributes (24 features + 1 outcome attribute). We use the *Rotation Forest* ensembling technique with *Alternating Decision Tree* as the underlying classifier to predict 5-year mortality. This technique generated a predictive model (c-statistic 0.86) that outperformed models generated using other techniques: logistic regression (c-statistic: 0.69), support vector machines (0.65), J48 decision trees (0.66), neural networks (0.73), naïve Bayes (0.80), random forest (0.80), and Bayesian networks (.82). Cross-validation was used to evaluate all the methods. The ensemble index was significantly better than all the above techniques in terms of c-statistic ($p<.001$ for all comparisons).

Logistic Regression[2] is used for prediction of the probability of occurrence of an event by fitting data to a sigmoidal S-shaped logistic curve. Logistic regression is often used with ridge estimators[3] to improve the parameter estimates and to reduce the error made by further predictions. *Support vector machines*[4] attempt to perform classification by constructing hyperplanes in a multidimensional space that separates the cases of different class labels. Different types of kernels can be used in SVM models, like linear, polynomial, radial basis function, and sigmoid. *J48* *algorithm*[5]is a decision tree approach with the internal nodes denoting the different attributes and the branches denoting the possible values of the attributes, while the leaf nodes indicate the final predicted value of the target variable. *Artificial neural networks* are networks of interconnected artificial neurons, and are commonly used for non-linear statistical data modeling to model complex relationships between inputs and outputs.[6,7] *The naive bayes classifier*[8] is a simple probabilistic classifier that is based upon the Bayes theorem. Although it makes strong assumptions about the independence of the input features it works well in practice for a wide variety of datasets and often outperforms other complex classifiers. *The Random Forest* [9] is an ensemble classifier that consists of multiple decision trees. The final class of an instance in a Random Forest is assigned by outputting the class that is the mode of the outputs of individual trees. *A Bayesian network* is a graphical model that encodes probabilistic relationships among a set of variables, representing a set of random variables and their conditional dependencies via a directed acyclic graph.[10]All predictive modeling was done using WEKA implementations of various techniques with default parameters, unless otherwise stated.

Rotation forest [11] is a method for generating classifier ensembles based on feature extraction, which can work both with classification and regression base learners. The training data for a the underlying classifier is created by applying Principal Component Analysis (PCA) [12] to *K* (here, *K*=10) subsets of the feature set, retaining all principal components in order to preserve the variability information in the data. Thus, *K* axis rotations take place to form the new features for the underlying classifier, to encourage simultaneously individual accuracy and diversity within the ensemble.

An *alternating decision tree* (ADTree) [13] is a machine learning technique that is a generalization of the classic decision tree algorithm. An alternating decision tree consists of two different types of nodes: decision nodes and prediction nodes. Decision nodes specify a predicate condition (like 'age' < 70). Prediction nodes contain a single real-value number. ADTrees always have prediction nodes as both root and leaves. An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing the values of any prediction nodes that are traversed. This is different from many other decision tree algorithms in which an instance follows only one path through the tree.

As an illustration, the alternating decision tree technique applied directly to the outcome prediction database results in the decision tree, partially shown in Figure 1. Of course, applying rotation forest ensembling technique provides a rotated attribute set to ADTree, which significantly improves model accuracy, but its visualization and interpretation is not straightforward and hence not shown here.

We further used the *Stacking* technique [14] with logistic modeling as the meta learning technique to calibrate the predictions from the advanced decision tree ensemble model. A new model is built on a new dataset, where the input is the predicted probabilities from the advanced decision tree ensemble model, and the output is the known outcome. To avoid over-fitting, 10-fold cross-validation is used to construct the new dataset (cross-validation is also used to evaluate overall model performance, as described later). A test instance with unknown outcome is thus classified by first passing it through the advanced decision tree ensemble model to get the intermediate probabilities, which are then passed through the meta-logistic model to get the final calibrated probabilities.

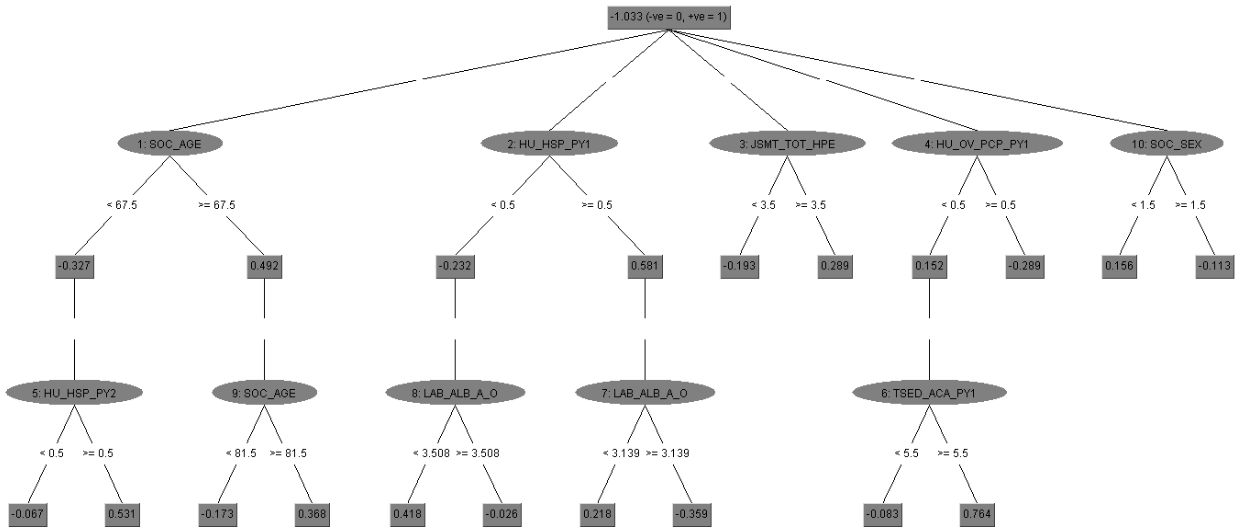

Figure 1 Alternating decision tree (partial illustration)

Model Evaluation

10-fold cross validation was used to evaluate the final model performance. The outcome prediction database was randomly divided into 10 segments with similar distribution of death and non-death cases as in the entire dataset; 9 segments were used for building the model and the remaining 1 segment was used to test the model. This procedure is repeated 10 times with different test segments. In this way, each instance in the dataset is tested exactly once using a model that did not see that instance while training. Binary classification performance can be evaluated using various metrics. We use the following in this work:

c-statistic (AUC): The ROC (Receiver operating characteristic) curve is a graphical plot of true positive rate and false positive rate. The area under the ROC curve (AUC or c-statistic) is an effective metric for evaluating binary classification performance, as it is independent of the probability cutoff and measures the discrimination power of the model.

Percentage of correct predictions: For highly unbalanced classes where the minority class is the class of interest, percentage of correct predictions by itself may not be a very useful indicator of classification performance, since even a trivial classifier which simply predicts the majority class would give a highpercentage of correct predictions.

$$\textit{Percentage of correct predictions} = (TP+TN)/(TP+TN+FP+FN)$$

where *TP* is the number of true positives (hits), *TN* is number of true negatives (correct rejections), *FP* is number of false positives (false alarms), and *FN* is number of false negatives (misses).

Sensitivity (Recall): It is the percentage of positive labeled records that were predicted positive. Recall measures the completeness of the positive predictions.

$$\textit{Sensitivity} = TP/(TP+FN)$$

Specificity: It is the percentage of negative labeled records that were predicted negative, thus measuring the completeness of the negative predictions.

$$\textit{Specificity} = TN/(TN+FP)$$

Positive predictive value (Precision): It is the percentage of positive predictions that are correct. Precision measures the correctness of positive predictions.

$$\textit{Positive predictive value} = TP/(TP+FP)$$

Negative predictive value: It is the percentage of negative predictions that are correct, thereby measuring the correctness of negative predictions.

$$\textit{Negative predictive value} = TN/(TN+FN)$$

F-measure: It is in general, possible to have either good precision or good recall, at the cost of the other, and F-measure combines the two measures in a single metric by taking the harmonic mean of precision and recall.

$$F\text{-}measure = 2.precision.recall/(precision+recall)$$

References

1. Hall MA. *Correlation-based feature selection for machine learning.* 1999; http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.4643. Accessed June 19, 2012.
2. Homer D., Lemeshow S..1989. Applied Logistic Regression. John Wiley and Sons, Inc.
3. Cessie, S. le, van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.
4. Vapnik V. N., The nature of statistical learning theory. Springer, 1995.
5. Quinlan R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
6. Bishop C., Neural Networks for Pattern Recognition. Oxford: University Press, 1995.
7. Fausett L., Fundamentals of Neural Networks. New York, Prentice Hall, 1994.
8. John G H, Langley P: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.
9. Breiman L (2001). Random Forests. Machine Learning. 45(1):5-32.
10. Cooper G, Herskovits E., (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 9(4):309-347.
11. Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2006;28(10):1619-1630.
12. Jolliffe IT. *Principal Component Analysis.* New York, NY: Springer; 2002.
13. Freund Y, Mason L. The alternating decision tree learning algorithm. *Proceedings of the Sixteenth International Conference on Machine Learning.* 1999:124-133.
14. Wolpert DH. Stacked Generalization. *Neural Networks.* 1992;5:241--259.