# Prediction of hospitalization due to heart diseases by supervised learning methods

Wuyang Dai[1], Theodora S. Brisimi[1], William G. Adams[2], Theofanie Mela[3], Venkatesh Saligrama[1], and Ioannis Ch. Paschalidis[1,*]

[*]Corresponding author
Mailing Address: Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02215, USA
Tel: (617) 353-0434 Fax: (617) 353-6440
E-mail: yannisp@bu.edu

[1]Department of Electrical & Computer Engineering, and Division of Systems Engineering, Boston University, 8 Saint Mary's St., Boston, MA 02215,
[2]Department of Pediatrics, Boston University School of Medicine and Boston Medical Center, 88 East Concord St, Boston, MA 02118, and
[3]Electrophysiology Lab/Arrhythmia Service, Massachusetts General Hospital, 55 Fruit St., Boston, MA 02114.

**Key words**: prevention, predictive models, hospitalization, heart diseases, machine learning, Electronic Health Records (EHRs).

**ABSTRACT**

**Background**. In 2008, the United States spent $2.2 trillion for healthcare, which was 15.5% of its GDP. 31% of this expenditure is attributed to hospital care. Evidently, even modest reductions in hospital care costs matter. A 2009 study showed that nearly $30.8 billion in hospital care cost during 2006 was potentially preventable, with heart diseases being responsible for about 31% of that amount.

**Methods.** Our goal is to accurately and efficiently predict heart-related hospitalizations based on the available patient-specific medical history. To the best of our knowledge, the approaches we introduce are novel for this problem. The prediction of hospitalization is formulated as a supervised classification problem. We use de-identified *Electronic Health Record (EHR)* data from a large urban hospital in Boston to identify patients with heart diseases. Patients are labeled and randomly partitioned into a training and a test set. We apply five machine learning algorithms, namely Support Vector Machines (SVM), AdaBoost using trees as the weak learner, Logistic Regression, a naïve Bayes event classifier, and a variation of a Likelihood Ratio Test adapted to the specific problem. Each model is trained on the training set and then tested on the test set.

**Results**. All five models show consistent results, which could, to some extent, indicate the limit of the achievable prediction accuracy. Our results show that with under 30% false alarm rate, the detection rate could be as high as 82%. These accuracy rates translate to a considerable amount of potential savings, if used in practice.

## 1. INTRODUCTION

The US health care system is considered costly and highly inefficient, devoting substantial resources to the treatment of acute conditions in a hospital setting rather than focusing on prevention and keeping patients out of the hospital. According to a recent study,[1] nearly $30.8 billion in hospital care cost during 2006 was preventable. Leading contributors were heart-related diseases accounting for more than $9 billion, or about 31%. Clearly, even modest percentage reductions in these amounts matter. This motivates our research to predict heart-related hospitalization. Two key enablers to such research are: the availability of patient EHRs and the existence of sophisticated (machine learning) algorithms that can process and learn from the data.

The adoption of EHRs into medical practices started more than two decades ago and EHRs have found diverse uses [20] e.g., in assisting the quality management in hospitals,[2] in detecting adverse drug reactions,[3] and in general primary care.[4] These early applications use EHRs for record keeping and information sharing and merely scratch the surface of what may be possible. Our belief is that the *true potential of EHRs* lies in their predictive ability of future acute health episodes and in guiding decision making. Foreseeing future hospitalizations for a large population of patients can drive preventive actions, such as scheduling a visit to the doctor, more frequent and exhaustive screening, calls by case nurses to assure medication adherence, or other mild interventions. All of these actions are much less costly than a hospitalization and, if successful, can drastically reduce hospital care costs. To that end, machine

learning methods seem to be promising tools and we extensively explore them for our problem.

Machine learning techniques have recently found use in various health-care applications. Vaithianathan et al.[5] uses multivariate logistic regression, a supervised learning method, to predict re-admissions in the 12 months following the date of discharge. Kim et al.[6] also employs two supervised learning algorithms and additionally incorporates interpretability of the models into consideration. This interpretability of results is also what we emphasis as an important criterion of method evaluation. Based on insurance claims data, Bertsimas et al.[7] combine spectral clustering (unsupervised method) with classification trees (supervised method) to first group similar patients into clusters and then make more accurate predictions about the near-future health-care cost. More closely related to our work are the prediction of re-admissions[8, 9] and the prediction of either death or hospitalization due to congestive heart failure.[10, 11] However, we differ from this line of work in that we do not limit our study to patients who are already admitted or to patients with a specific heart ailment. This makes our setting novel and broader.

Our algorithms consider the history of a patient's records and predict whether each individual patient will be hospitalized in the following year, thereby, alerting the health care system and potentially triggering preventive actions. An obvious advantage of our algorithmic approach is that it can easily scale to a very large number of monitored patients; such scale is not possible with human monitors. Our results

suggest that with about 30% false positives, 82% of heart-related hospitalizations can be accurately predicted. An important contribution is that these accuracy rates surpass what is possible with more empirical but well accepted risk metrics, such as a heart disease risk factor that emerged out of the Framingham study.[12] We show that even a more sophisticated use of the features used in the Framingham risk factor, still leads to results inferior to our approaches. This suggests that the *entirety of a patient's EHR is useful in the prediction and this can only be achieved with a systematic algorithmic approach.*

The remainder of the paper is organized as follows. Section 2 contains a detailed description of the data set, the preprocessing steps, the methods we propose for hospitalization prediction, and the criteria we apply for evaluating the performance of the methods. Section 3 contains our experimental results. A discussion of the results is in Section 4. We end with some concluding remarks in Section 5.

## 2. DATA AND METHODS

### 2.1 Detailed Data Description and Objective

The data we used are from the Boston Medical Center (BMC) – the largest safety-net hospital in New England. The study is focused on patients with at least one heart-related diagnosis or procedure record in the period 01/01/2005–12/31/2010. For each patient in the above set, we extract the medical history (demographics, visit history, problems, medications, labs, procedures and limited clinical observations) for the period 01/01/2001–12/31/2010, which contains relevant **medical factors** and

from which the features of the dataset will be formed. Data were available from the hospital EHR and billing systems (which record admissions or visits and the primary diagnosis/reason). The various categories of medical factors, along with the number of factors and some examples corresponding to each, are shown in Table 1. We note that some of the Diagnoses and Admissions are not directly heart-related, but may be good indicators of a heart problem. Overall, our data set contains 45,579 patients. 60% of that set forms our *training set* – used for training algorithms – and the remaining 40% is designated as the *test set* and used exclusive for evaluating the performance of the algorithms.

Our objective is to leverage past medical factors for each patient to predict whether she/he will be hospitalized or not during a **target** year which could be different for each patient.

**Table I. Medical Factors.**

| **Category** | **Number of Factors** | **Examples** |
|---|---|---|
| Demographics | 4 | Sex, Age, Race, Zip Code |
| Diagnoses | 22 | e.g., Acute Myocardial Infarction (ICD9: 410), Cardiac Dysrhythmias (ICD9: 427), Heart Failure (ICD9: 428), Acute Pulmonary Heart Disease (ICD9: 415), Diabetes Mellitus with Complications (ICD9: 250.1-250.4, 250.6-250.9), Obesity (ICD9: 278.0) |
| Procedures CPT | 3 | Cardiovascular Procedures (including CPT 93501, 93503, 93505, etc.), Surgical Procedures on the Arteries and Vein (including CPT |

| | | |
|---|---|---|
| | | 35686, 35501, 35509, etc.), Surgical Procedures on the Heart and Pericardium (including CPT 33533, 33534, 33535) |
| Procedures ICD9 | 4 | Operations on the Cardiovascular System (ICD9: 35-39.99), Cardiac Stress Test and pacemaker checks (ICD9: 89.4), Angiocardiography and Aortography (ICD9: 88.5), Diagnostic Ultrasound of Heart (ICD9: 88.72) |
| Vitals | 2 | Diastolic Blood Pressure, Systolic Blood Pressure |
| Lab Tests | 4 | CPK (Creatine phosphokinase) (LOINC:2157-6), CRP Cardio (C-reactive protein) (LOINC:30522-7), Direct LDL (Low-density lipoprotein) (LOINC:2574-2), HDL (High-density lipoprotein) (LOINC:9830-1) |
| Tobacco | 2 | Current Cigarette Use, Ever Cigarette Use |
| Visits to the Emergency Room | 1 | Visits to the Emergency Room |
| Admissions | 17 | e.g., Heart Transplant or Implant of Heart Assist System (MSDRG: 001, 002), Cardiac Valve and Other Major Cardiothoracic procedures (MSDRG: 216-221), Coronary Bypass(MSDRG: 231-234), Acute Myocardial Infarction (MSDRG: 280-285), Heart Failure and Shock (MSDRG: 291-293), Cardiac Arrest (MSDRG: 296-298), Chest Pain (MSDRG: 313), Respiratory System related admissions (MSDRG: 175-176, 190-192) |

In order to organize all the available information in some uniform way for all patients, some preprocessing of the data is needed to summarize the information over a time interval. Details will be discussed in the next subsection. We will refer to the summarized information of the medical factors over a specific time interval as **features**.

Each feature related to Diagnoses, Procedures CPT, Procedures ICD9 and Visits to the Emergency Room is an integer count of such records for a specific patient during the specific time interval. Zero indicates absence of any record. Blood pressure and lab tests features are continuous-valued. Missing values are replaced by the average of values of patients with a record at the same time interval. Features related to tobacco use are indicators of current- or past-smoker in the specific time interval. Admission features contain the total number of days of hospitalization over the specific time interval the feature corresponds to. Admission records are used both to form the Admission features (past admission records) and in order to calculate the prediction variable (existence of admission records in the target year). We treat our problem as a classification problem and each patient is assigned a **label**: 1 if there is a heart-related hospitalization in the target year and 0 otherwise.

### 2.2 Data Preprocessing

In this subsection we discuss several data organization and preprocessing choices we make. For each patient, a target year is fixed (the year in which a hospitalization prediction is sought) and all past patient records are organized as follows.

- *Summarization of the medical factors in the history of a patient*: Based on experimentation, an effective way to summarize each patient's medical history is to form four time blocks for each medical factor with all corresponding records summarized over one, two, and three years before the target year and all

earlier records being summarized in a fourth block. For blood pressure and tobacco use, only the year before the target year is kept. This process results to a vector of 212 features for each patient.

- *Selection of the target year*: As a result of the nature of the data, the two classes are highly imbalanced. When we fix the target year for all patients to be 2010, the number of hospitalized patients is about 2% of the total number of patients, which makes the classification problem much more challenging. Thus, and to increase the number of hospitalized patient examples, if a patient had only one hospitalization throughout 2007-2010, the year of hospitalization is set as the target year for that patient. If a patient had multiple hospitalizations, a target year between the first and the last hospitalization is randomly selected.
- *Setting the target time interval to be a year*: A year has been proven to be an appropriate time interval for prediction for our data set. We conducted trials setting the time interval for prediction to be 1, 3, 6 and 12 months and used a Support Vector Machine classifier — a method described later in more detail. Setting the target time interval to one year yielded the best results. Moreover, given that hospitalization occurs roughly uniformly within a year, we take the prediction time interval to be a calendar year.
- *Removing noisy samples*: Patients who have no records before the target year are impossible to predict and are thus removed.

After preprocessing, the samples are labeled as belonging to the hospitalized or non-hospitalized class. The ratio between the two classes is 14:1, which is highly imbalanced. More specifically, the number of patients from the hospitalized class in our dataset is 3,033 which is large enough to accommodate sufficient training and testing. This imbalance prevents us to later report a single classification error number, because one class would dominate the other. Instead, we consider two types of performance rates separately, namely, false alarm rates and detection rates, which are presented later in detail. It is also worth mentioning that this disproportion of the two classes also affects the design of our new algorithm (*K*-LRT) described in the next section.

### 2.3 Proposed Methods

To predict whether patients are going to be hospitalized in the target year given their medical history, we experiment with five different methods. All five are typical examples of supervised machine learning. We adapt the last one to better fit the specific application we examine. The first three methods fall into the category of *discriminative learning algorithms*, while the latter two are *generative algorithms*. Discriminative algorithms directly partition the input space into label regions without modeling how the data are generated, while generative algorithms assume a model that generates the data, estimate the model's parameters and use it to make classifications. Discriminative methods are likely to give higher accuracy, but generative methods provide more interpretable models and results. This is the reason we experiment with

methods from both families and the trade-off between accuracy and interpretability is observed in our results.

**2.3.1 Support Vector Machines (SVM)**. An SVM is a very efficient two-category classifier.[13] Intuitively, the SVM algorithm attempts to find a separating hyperplane in the feature space, so that data points from different classes reside on different sides of that hyperplane. We can calculate the distance of each input data point from the hyperplane. The minimum over all these distances is called *margin*. The goal of SVM is to find the hyperplane that has the maximum margin. In many cases data points are neither linearly nor perfectly separable. To that end, one can make the classifier tolerant to some misclassification errors and leverage kernel functions to "elevate" the features into a higher dimensional space where linear separability is possible.[13] We employ the widely used Radial Basis Function (RBF)[14] as the kernel function in our experiment settings. Tuning parameters are the misclassification penalty coefficient and the kernel parameter; we used the values [0.3, 1, 3] and [0.5, 1, 2, 7 15, 25, 35, 50, 70, 100], respectively. Optimal values of 1 and 7, respectively, were selected by cross-validation.

**2.3.2 AdaBoost with Trees.** Boosting[15] provides an effective way of combining decisions of not necessarily strong classifiers to produce highly accurate predictions. The AdaBoost algorithm iteratively adjusts the weights of various training data points through an exponential up-weighting or down-weighting

procedure. Specifically, starting with equal weights, the algorithm generates in every iteration a new base classifier to best fit the current weighted samples. Then, the weights are updated so that the misclassified samples are assigned higher weights so as to influence the training of the next base classifier. At termination, a weighted combination of the base classifiers is the output of AdaBoost. In our study we use *stumps*, which are two-level *Classification and Regression Trees (CART)*, as the base classifier.[16] This method recursively partitions the space into a set of rectangles and then fits a prediction within each partition.

There is an extra preprocessing step applied to the data. The zip code values are clustered into 4 clusters using the k-means algorithm[16] and this feature is treated as a categorical one. The number of iterations in the Adaboost method is a model parameter which can be tuned by cross-validation. In our case, this tuning led to setting to 100,000 the number of Adaboost iterations.

**2.3.3 Logistic Regression.** Logistic Regression[19] is a popular classification method used in many applications. This method models the posterior probability that a sample falls into a certain class (e.g., the positive class) as a logistic function and the input of this logistic function is the linear combination of the input features. Under this model, the log-likelihood ratio of the posterior probabilities of the two classes is a linear function of the input features. Therefore, the decision boundary that separates the two classes is still linear. However,

beyond the classification decision, the prediction on a certain sample point naturally comes with a probability value, which could be meaningful in many applications. Thus, logistic regression is widely used.

**2.3.4 Naïve Bayes Event Model**. Naïve Bayes models are generative models that assume the features or "events" to be generated independently (naïve Bayes assumption[17]). Naïve Bayes classifiers are among the simplest models in machine learning, but despite their simplicity, they work quite well in real applications. There are two types of naïve Bayes models.[17] The first one will be presented extensively in the next method. The second one, referred to as the *Naïve Bayes Event Model*, works as follows. To generate a new patient from the model, a label $y$ will first be generated (hospitalized or non-hospitalized) based on a prior distribution $p(y)$. Then, for this patient, a sequence of events ($x_t$'s) is generated by choosing each event independently from certain multinomial conditional distributions $p(x|y)$. An event can appear many times for a patient and the overall probability of this newly generated patient is the product of the class prior with the product of the probabilities of each event. In our problem, an event is a specific combination of the medical factors. We consider only the medical factors from the following six categories: Diagnoses, Admissions, Emergency, Procedures CPT, Procedures ICD9, and Lab Tests. To generate such a data set, we aggregate the medical factors that belong to each one of these types and count the total number of records of the same type in each of the four time blocks discussed earlier that represent a patient's history.

Thus, each patient is represented as a sequence of four events. To make events more intuitive and to reduce the total number of possible events, the data just formed are quantized into binary values and then the tuples of the six binary values (one for each category) are encoded into $2^6$ single values. We estimate the prior distribution of labels $p(y)$ and the conditional distributions $p(x|y)$ from the training set and make predictions for the test set based on the likelihoods calculated from these distributions.

**2.3.5 *K*-Likelihood Ratio Test.** The *Likelihood Ratio Test (LRT)* is a Naïve Bayes classifier and, as described before, assumes that features $x_i$ are independent. For this method as well, we quantize the data as shown in Table 2.

**Table II. Quantization of Features.**

| Features | Levels of quantization | Comments |
|---|---|---|
| Sex | 3 | 0 represents missing information |
| Age | 6 | Thresholds at 40, 55, 65, 75 and 85 years old |
| Race | 10 | |
| Zip Code | 0 | Removed due to its vast variation |
| Tobacco (Current and Ever Cigarette Use) | 2 | Indicators of cigarette use |
| Diastolic Blood Pressure (DBP) | 3 | Level 1 if DBP < 60mmHg, Level 2 if 60mmHg ≤ DBP ≤ 90mmHg and Level 3 if DBP > 90mmHg |

| | | |
|---|---|---|
| Systolic Blood Pressure (SBP) | 3 | Level 1 if SBP < 90mmHg, Level 2 if 90mmHg ≤ SBP ≤ 140mmHg and Level 3 if SBP > 140mmHg |
| Lab Tests | 2 | Existing lab record or Non-Existing lab record in the specific time period |
| All other dimensions | 7 | Thresholds are set to 0.01%, 5%, 10%, 20%, 40% and 70% of the maximum value of each dimension |

In the quantized data set, the LRT algorithm (see also [21]) empirically estimates the distribution $p(x_i|y)$ of each feature for the hospitalized and the non-hospitalized class. Given a new test sample $\boldsymbol{z}=\{z_1, z_2, \dots, z_n\}$, LRT calculates the two likelihoods $p(z|y=1)$ and $p(z|y=0)$ ($y=0$ corresponds to non-hospitalized and $y=1$ to hospitalized) and then classifies the sample based on the ratio $p(z|y=1)/\ p(z|y=0)$. Due to independence, the ratio $p(z|y=1)/\ p(z|y=0)$ is the product of $p(z_i|y=1)/\ p(z_i|y=0)$ over i. In our variation of the method, which we will call *K*-LRT, instead of taking into account the ratios of the likelihoods of all features, we consider only the K features with the largest ratios. This type of method is closely related to the anomaly detection methods in [18] and [22]. The purpose of this "feature selection" is to identify the *K most significant* features for each individual patient. Thus, each patient is actually treated differently. After experimentation, the best performance is achieved by setting K=4. The prediction accuracy for K=1 is also reported in the experimental results section. It is worth mentioning that the K most significant features are with

respect to the hospitalized class. We deliberately chose this unbalanced strategy (tilting towards the hospitalized class) mainly because of two reasons. The first one is that the sample size of the hospitalized class is much smaller than the non-hospitalized class (1:14). As a result, a strong non-hospitalized signal (i.e., a small value of $p(z_i|y=1)/\ p(z_i|y=0)$ for some feature) could simply be due to underestimating the tail of the distribution of feature i for the hospitalized class. The second reason is actually drawn from the results in Figure 2. The accuracies of 1-LRT, 4-LRT and LRT (the latter using all features) are almost the same, which validates the proposed method.

### 2.4 Evaluation Criteria

Typically, the primary goal of learning algorithms is to maximize the prediction accuracy or equivalently minimize the error rate. However, in the specific medical application problem we study, the ultimate goal is to alert and assist doctors in taking further actions to prevent hospitalizations before they occur, whenever possible. Thus, our models and results should be accessible and easily explainable to doctors and not only machine learning experts. With that in mind, we examine our models from two aspects: prediction accuracy and interpretability.

#### 2.4.1 Prediction Accuracy

The prediction accuracy is captured in two metrics: the *False Alarm Rate* (the fraction of false positives out of the negatives) and the *Detection Rate* (the fraction of true positives out of the positives). Note that in the medical literature, the detection rate is

often referred to as *sensitivity* and the term *specificity* is used for one minus the false alarm rate. For a binary classification system, the evaluation of the performance using these two metrics is typically illustrated with the *Receiver Operating Characteristic (ROC)* curve, which plots the Detection Rate versus the False Alarm Rate at various threshold settings.

### 2.4.2 Interpretability

With SVM, the features are mapped through a kernel function from the original space into a higher-dimensional space. This, however, makes the features in the new space not interpretable. In AdaBoost with trees, while a single tree classifier which is used as the base learner is explainable, the weighted sum of a large number of trees makes it relatively complicated to find the direct attribution of each feature to the final decision. The naïve Bayes Event model is in general interpretable, but in our specific problem each patient has a relatively small sequence of events (four) and each event is a composition of medical factors. Thus, again, to find the direct attribution of each feature to the final decision is hard. LRT itself and Logistic Regression still lack interpretability, because we have more than 200 features for each sample and there is no direct relationship between prediction of hospitalization and the reasons that led to it. The most interpretable method is *K*-LRT. *K*-LRT highlights the top *K* features that lead to the classification decision. These features could be of help in assisting the physicians reviewing the patient's EHR profile.

# 3 RESULTS

### 3.1 Insight into the Data

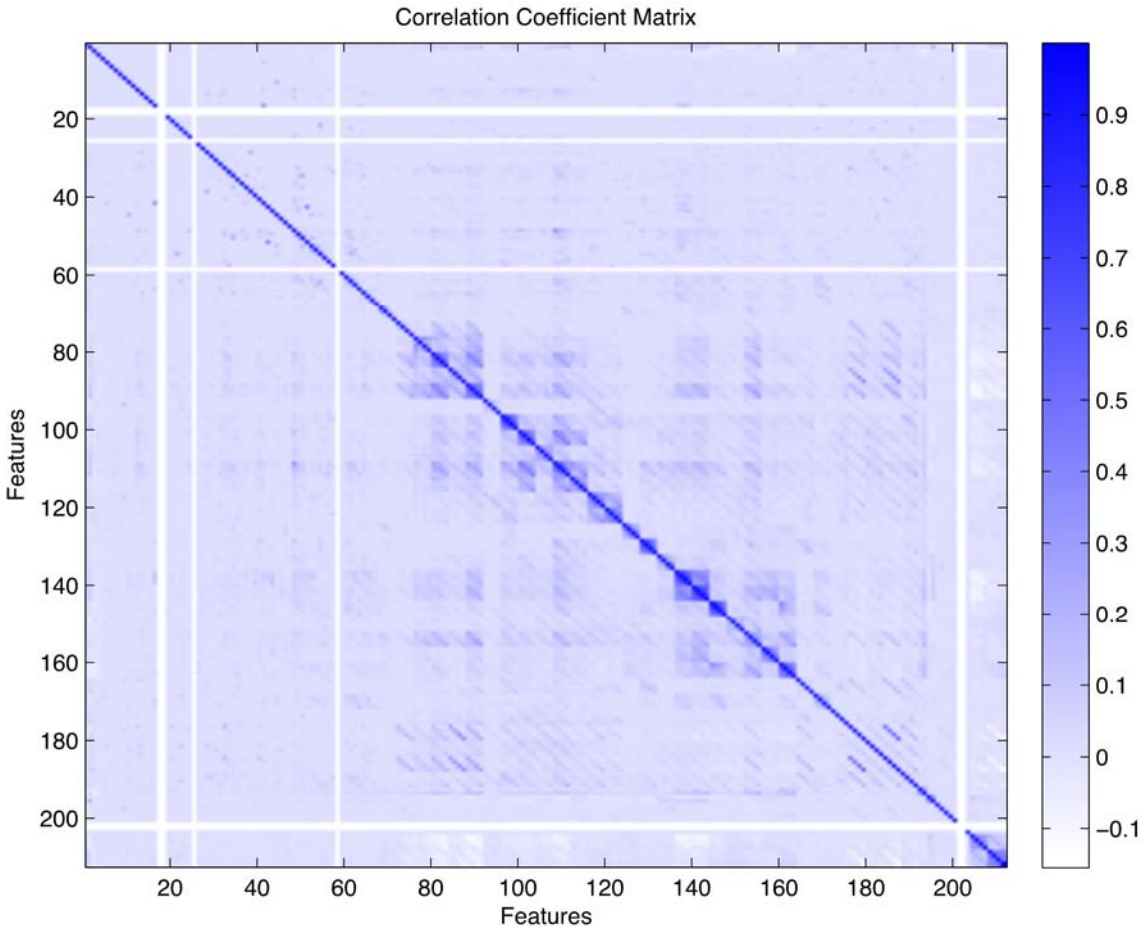


**Figure 1. Each point (i,j) corresponds to the correlation coefficient between Feature i and Feature j. There are a few features with zero variance (shown as white stripes) that are later removed from the feature set.**

The correlation coefficient matrix of all features is shown in Figure 1. Most of the features are weakly correlated. There is moderate correlation between features that refer to the same medical factor but correspond to different time blocks (near-diagonal elements) and between few other pairs of features including: Diagnosis of Chronic Ischemic Heart Disease with Diagnosis of Diabetes, Diagnosis of Ischemic Heart Dis-

ease with Diagnosis of Old Myocardial Infarction, Diagnosis of Heart Failure with Admission due to Heart Failure, and Operations on Cardiovascular System with Ultrasound of the Heart.

### 3.2 Prediction Accuracy results

We first compare the performance of LRT using all features and *K*-LRT under different values of *K*. Figure 2 shows the prediction accuracy for LRT, *1*-LRT and *4*-LRT. In Figure 3, a comparison of the performance of all five methods we presented is illustrated. We also generate the ROC curve based on patients' 10-year risk of General Cardiovascular Disease defined in the Framingham Heart Study (FHS).[12] FHS is a seminal study on heart diseases that has developed a set of risk factors for various heart problems. The 10-years risk we are using is the closest to our purpose and has been widely used. We calculate this risk value (which we call the *Framingham Risk Factor-FRF*) for every patient and make the classification based on this risk factor only. We also generate an ROC by applying the AdaBoost with trees method just to the features involved in FRF. The generated ROC serves as a baseline for comparison.

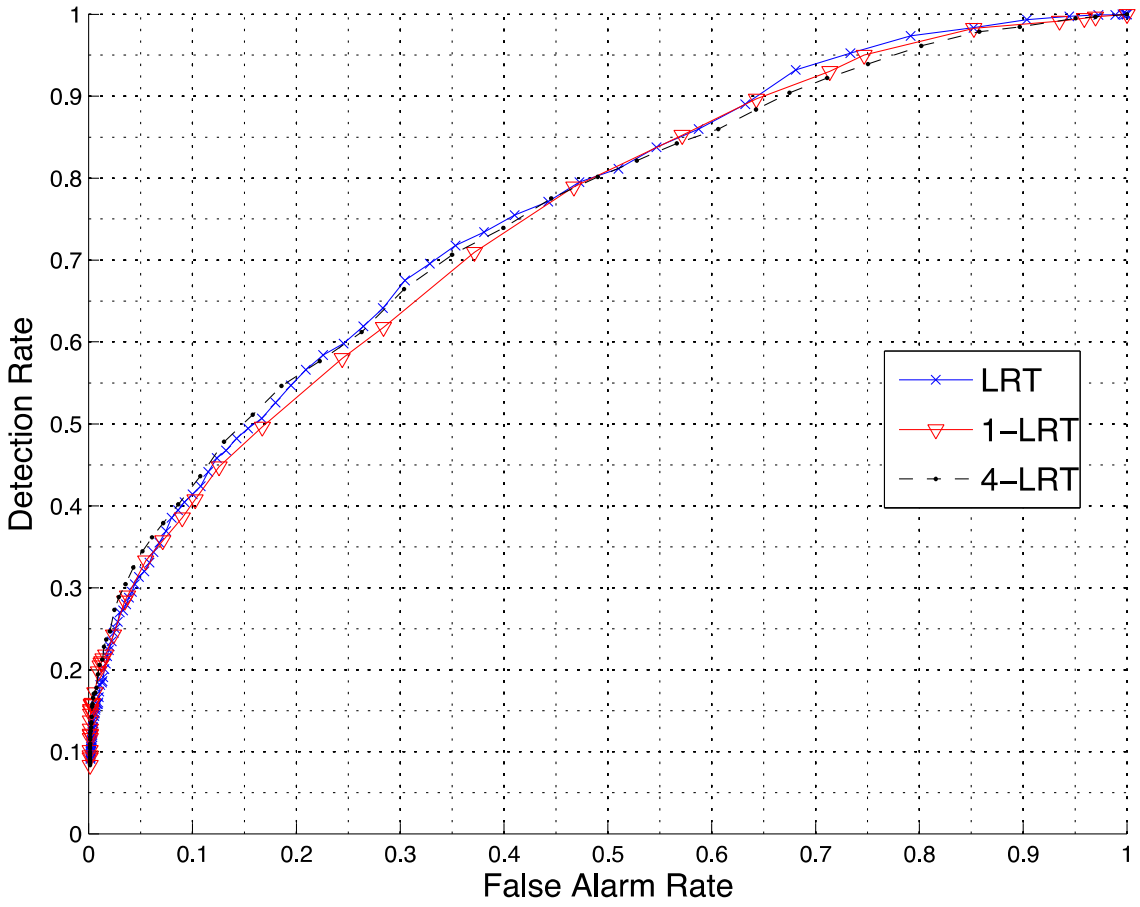


**Figure 2.** Comparison of LRT, *1*-LRT and *4*-LRT.

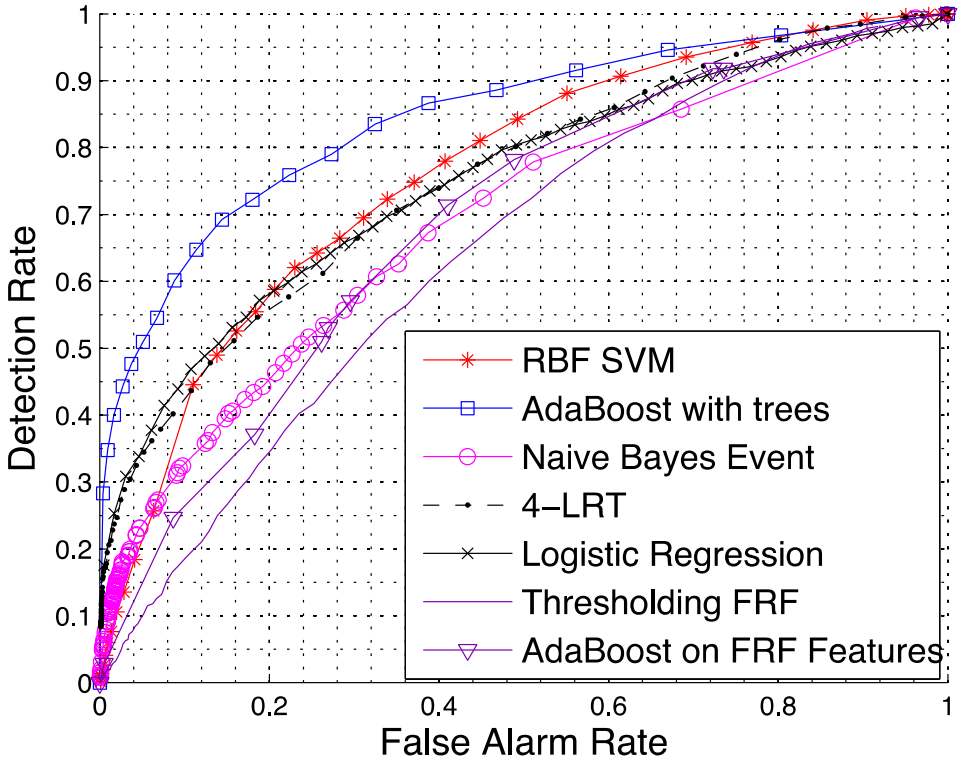


**Figure 3.** Comparison of all five methods and the methods based on the Framingham Heart Study.

### 3.3 Interpretability results

Below we present the features highlighted by two of our methods: *1*-LRT and Adaboost. We remind the reader that in *1*-LRT, each test patient is essentially associated with a single feature. For all features, we count how many times they were selected as the primary feature and we report in Table 3 (left column) the 10 features that were the most popular as primary. Adaboost, on the other hand, yields a linear combination of decision trees and is hard to interpret. However, we can calculate a variable *Importance Score (IS)*[1][16] for each feature, which highlights the most significant

[1] The 'Importance score' in Adaboost is calculated by summing changes in the risk (difference between the risk for the parent node and the total risk for the two children) due to splits on every predictor and dividing the sum by the number of branch nodes. The sum is taken over the best splits

features. Table 3 also lists the top 10 important features indicated by their importance score (right column). Features that appear in both columns are in bold.

**Table III. Top 10 significant features for 1-LRT and AdaBoost with Trees.**

| **1-LRT** | | **AdaBoost with Trees** | |
|---|---|---|---|
| **Counts** | **Feature Name** | **IS** ($x10^{-4}$) | **Feature Name** |
| 1591 | ***Age*** | 0.6462 | *Diagnosis of diabetes mellitus w/o complications, 1 year before the target year* |
| 548 | ***Visit to the Emergency Room, 1 year before the target year*** | 0.5498 | ***Diagnosis of heart failure, 1 year before the target year*** |
| 525 | Diagnosis of hematologic disease, 1 year before the target year | 0.4139 | ***Age*** |
| 523 | ***Diagnosis of heart failure, 1 year before the target year*** | 0.3187 | ***Symptoms involving respiratory system and other chest symptoms, 1 year before the target year*** |
| 514 | ***Symptoms involving respiratory system and other chest symptoms, 1 year before the target year*** | 0.2470 | Admission due to other circulatory system diagnoses, 1 year before the target year |

found at each branch node. The risk of a node is the node error weighted by the probability of that node.

| | | | |
|---|---|---|---|
| 486 | ***Diagnosis of diabetes mellitus w/o complications, 1 year before the target year*** | 0.2240 | Visit to the Emergency Room, 4 years before the target year and the rest of the history |
| 474 | Lab test CPK, 1 year before the target year | 0.1957 | Operations on cardiovascular system (heart and septa OR vessels of heart OR heart and pericardium), 4 years before the target year and the rest of the history |
| 451 | Lab test CPK, 4 years before the target year and the rest of the history | 0.1578 | ***Visit to the Emergency Room, 1 year before the target year*** |
| 408 | ***Diagnosis of heart failure, 2 years before the target year*** | 0.1543 | Symptoms involving respiratory system and other chest symptoms, 4 years before the target year and the rest of the history |
| 356 | Diagnosis of diabetes mellitus w/o complications, 2 years before the target year | 0.1124 | ***Diagnosis of heart failure, 2 year before the target year*** |

To provide additional insight into the algorithms, Table 4 presents five more medically significant features highlighted by each method and two interesting features with low significance in both methods. For *1*-LRT, features with low significance are the ones with a likelihood ratio $p(z_i|y=1)/\ p(z_i|y=0)$ close to 1. For Adaboost, non-significant features have a low IS.

**Table IV. Other significant and non-significant features with 1-LRT and AdaBoost with Trees.**

| **Another 5 significant features in 1-LRT** | **Another 5 significant features in AdaBoost with Trees** |
|---|---|
| Lab Test High-density lipoprotein (HDL) | Lab Test High-density lipoprotein (HDL), 1 year before the target year |
| Lab Test Low-density lipoprotein (LDL) | Angiography and Aortography procedures, 4 years before the target year and the rest of the history |
| Systolic Blood Pressure | Cardiac Catheterization Procedures, 4 years before the target year and the rest of the history |
| Diagnosis of Heart Failure | Race |
| Diagnosis of Other Forms of Chronic Ischemic Heart Diseases | Cardiac Dysrhythmias, 1 year before the target year |
| **2 non-significant features in 1-LRT** | **2 non-significant features in AdaBoost with Trees** |
| Sex | Sex |
| Hypertensive Heart Disease, 1 year before the target year | Hypertensive Heart Disease, 1 year before the target year |

## 4 DISCUSSION

Based on the experimental results regarding the accuracy of our methods (Section 3.1), we draw the following conclusions:

1. LRT, *1*-LRT and *4*-LRT achieve very similar performance (the corresponding ROC curves of the three methods are close to each other). This indicates that

using only the most significant or several significant features with the largest likelihood ratios, is sufficient in making an accurate prediction. It also suggests that our problem is close to an "anomaly detection" problem and identifying the most anomalous feature captures most of the information that is useful for classification.

2. From the comparison of all five methods in Figure 3, it can be seen that AdaBoost is the most powerful one and performs the best except for situations that require very low false alarm rates. Put it differently, if we fix the false alarm rate, AdaBoost achieves the highest detection rate among all methods, and conversely, if we fix the detection rate, AdaBoost yields the lowest false alarm rate. On the other hand, the Naïve Bayes Event classifier generally performs the worst due to its simplicity.

3. The performance of RBF SVM, Logistic Regression, AdaBoost with trees, and *4*-LRT is quite similar in general (the corresponding ROC curves do not differ much). However, these methods have very different assumptions and underlying mathematical formulation. Based on this observation, we conjecture that we have approached the limit of the prediction accuracy that could be achieved with the available data.

4. All of our proposed methods perform better than utilizing the FRF, except for the naïve Bayes event classifier for high false alarms rates (i.e., the ROC curves that correspond to FRF features are worse in the sense described above compared to the rest of the methods). Even applying AdaBoost with trees (the best method so far) to the features involved in calculating the FRF, does not seem to

help a lot. This suggests that it is valuable to have and leverage a multitude of patient-specific features obtained from EHRs. Using these data, however, necessitates the use of the algorithmic approach we advocate.

Based on the results in Table III, it is clear that the two sets of features highlighted by the two methods have several features in common, indicating that the results from the different methods are consistent. This consistency supports the validity of our methods from a stability/sensitivity perspective as well.

From a medical point of view, the features listed in Table III are reasonably highlighted. ER visits, a diagnosis of heart failure, and chest pain or other respiratory symptoms are often pre-cursors of a major heart episode. The CPK test is also viewed as one of the most important tests for diagnosing Acute Myocardial Infarction (AMI) and AMI, among all heart diseases, is the most probable to lead to hospitalization.

What is interesting to note in Table V is that Hypertensive heart disease is considered non-significant by both methods. This is probably due to the fact that, once diagnosed, it is usually well-treated and the patient's blood pressure is well-controlled.

## 5 CONCLUSIONS

Our research is a novel attempt to predict hospitalization due to heart disease using various machine learning techniques. Our results show that with a 30% false alarm

rate, we can successfully predict 82% of the patients with heart diseases that are going to be hospitalized in the following year. We examine methods that have high prediction accuracy (Adaboost with trees), as well as methods that can help doctors identify features to help them when examining patients (K-LRT). One could choose which one to use depending on the ultimate goal and the desirable target for detection and false alarm rates. If coupled with case management and preventive interventions, our methods have the potential to prevent a significant number of hospitalizations by identifying patients at greatest risk and enhancing their outpatient care *before* they are hospitalized. This can lead to better patient care, but also to potentially substantial health care cost savings. In particular, even if a small fraction of the $30.8 billion spent annually on preventable hospitalizations can be realized in savings, this would offer significant benefits. Our methods also produce a set of significant features of the patients that lead to hospitalization. Most of these features are well-known precursors of heart problems, a fact which highlights the validity of our models and analysis. The methods are general enough and can easily handle new predictive variables as they become available in EHRs, to refine and potentially improve the accuracy of our predictions. Furthermore, methods of this type can also be used in related problems such as predicting re-hospitalizations.

## AUTHORS' CONTRIBUTIONS

W.D. and T.S.B. co-designed the methods, performed the analysis, produced results and figures, and co-wrote the manuscript. W.G.A. provided access to the data, advised on the use and interpretation of the data, provided medical intuition and commented

on the manuscript. T.M. advised on heart-related diseases, suggested relevant medical features, advised on the interpretation of the results, and commented on the manuscript. V.S. co-led the study, co-designed the methods, and commented on the manuscript. I.Ch.P. led the study, co-designed the methods, and co-wrote the manuscript.

## ACKNOWLEDGEMENTS

We would like to thank Dimitris Bertsimas and John Silberholz for useful discussions and suggestions.

This research has been partially supported by the NSF under grants IIS-1237022 and CNS-1239021, by the NIH/NIGMS under grant GM093147, by the ARO under grants W911NF-11-1-0227 and W911NF-12-1-0390, and by the ONR under grant N00014-10-1-0952.

## REFERENCES

1 Jiang J, Russo A, Barrett M. Nationwide frequency and costs of potentially preventable hospitalizations, 2006. *HCUP Statistical Brief 72,* April 2009, Agency for Healthcare Research and Quality, Rockville, MD. Available from: www.hcup-us.ahrq.gov/reports/statbriefs/sb72.pdf

2 Takeda H, Matsumura Y, Nakajima K, et al. Health care management by means of an incident report system and an electronic health patient record system. *International Journal of Medical Informatics* 2003;69:285-293.

3 Hannan J. Detecting adverse drug reactions to improve patient outcomes. *International Journal of Medical Informatics*. 1999;55:61-64.

4 Wang S, Middleton B, Prosser L, *et al*. A cost-benefit analysis of electronic medical records in primary care. *The American journal of medicine* 2003;**114**:397-403.

5 Vaithianathan R, Jiang N, Ashton T. A model for predicting readmission risk in New Zealand. *Working paper number 2012-02*, 2012.

6 Cho I, Park I, Kim E, *et al*. Using EHR data to predict hospital-acquired pressure ulcers: A prospective study of a Bayesian Network model. *International Journal of Medical Informatics.* 2013;82:1059-1067.

7 Bertsimas D, Bjarnadottir M V, Kane M A, *et al*. Algorithmic prediction of health-care costs. *Operations Research* 2008;**56**:1382-1392.

8 Agarwal J. *Predicting risk of re-hospitalization for congestive heart failure patients [MS Thesis]*. Seattle, WA: University of Washington; 2012.

9 Giamouzis G, Kalogeropoulos A, Georgiopoulou V, *et al*. Hospitalization epidemic in patients with heart failure: risk factors, risk prediction, knowledge gaps, and future directions. *Journal of cardiac failure* 2011;**17**:54-75.

10 Smith D, Johnson E, Thorp M, *et al*. Predicting Poor Outcomes in Heart Failure. *The Permanente Journal* 2011;**15**:4-11.

11 Wang L, Porter B, Maynard C, *et al*. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. *Medical Care* 2013;**51**:368-373.

12 D'Agostino R, Vasan R, Pencina M, *et al*. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;**117**:743-753.

13 Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995;**20**:273-297.

14 Schölkopf B, Sung K, Burges C, *et al*. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Sign. Processing* 1997;**45**:2758–2765.

15 Yoav F, Schapire R, Abe N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* 1999;**14**:771-780.

16 Hastie T, Tibshirani R, Friedmand J. *The Elements of Statistical Learning.* New York, NY: Springer 2009.

17 McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. *AAAI-98 workshop on learning for text categorization* 1998;**752**:41-48.

18 V. Saligrama, M. Zhao. Local Anomaly Detection. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS), April 21-23, 2012: 969-983. La Palma, Canary Islands.

19 Bishop, Christopher M., and Nasser M. Nasrabadi. *Pattern recognition and machine learning.* New York, NY: Springer, 2006.

20 Ludwick, Dave A., and John Doucette. "Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries." International journal of medical informatics 78.1 (2009): 22-31.

21 Paschalidis, Ioannis Ch, and Dong Guo. "Robust and distributed stochastic localization in sensor networks: Theory and experimental results." *ACM Transactions on Sensor Networks (TOSN)* 5.4 (2009): 34.

22 Paschalidis, Ioannis Ch, and Georgios Smaragdakis. "Spatio-temporal network anomaly detection by assessing deviations of empirical measures." *IEEE/ACM Transactions on Networking (TON)* 17.3 (2009): 685-697.