Early Prediction of Cardiac Arrest (Code Blue) using Electronic Medical Records

Sriram Somanchi
Carnegie Mellon University
Pittsburgh, USA
somanchi@cmu.edu

Samrachana Adhikari
Carnegie Mellon University
Pittsburgh, USA
asamrach@andrew.cmu.edu

Allen Lin
Harvard University
Cambridge, MA
allenlin@g.harvard.edu

Elena Eneva
Accenture
San Francisco, USA
elena.eneva@accenture.com

Rayid Ghani
University of Chicago
Chicago, USA
rayid@uchicago.edu

ABSTRACT

Code Blue is an emergency code used in hospitals to indicate when a patient goes into cardiac arrest and needs resuscitation. When Code Blue is called, an on-call medical team staffed by physicians and nurses is paged and rushes in to try to save the patient's life. It is an intense, chaotic, and resource-intensive process, and despite the considerable effort, survival rates are still less than 20% [4]. Research indicates that patients actually start showing clinical signs of deterioration some time before going into cardiac arrest[1][2][3], making early prediction, and possibly intervention, feasible. In this paper, we describe our work, in partnership with NorthShore University HealthSystem, that preemptively flags patients who are likely to go into cardiac arrest, using signals extracted from demographic information, hospitalization history, vitals and laboratory measurements in patient-level electronic medical records. We find that early prediction of Code Blue is possible and when compared with state of the art existing method used by hospitals (MEWS - Modified Early Warning Score) [4], our methods perform significantly better. Based on these results, this system is now being considered for deployment in hospital settings.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and science*

Keywords

Machine Learning; Data Mining; SVM; Code Blue; Cardiac Arrest; Electronic Medical Records; Early Prediction

KDD'15, August 11-14, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2788588.

1. INTRODUCTION

Code Blue is used to alert a rapid response team (RRT) in a hospital when a patient goes into cardiac arrest and needs resuscitation. The rapid response team intervenes, dropping other tasks they are attending to at the time, and rushes to the Code Blue patient to possibly improve his or her deteriorating condition. More than 200,000 adult in-hospital cardiac arrests occur in the United States each year, and as many as 80% of these incidents result in the death of the patient [4]. Medical professionals believe that if Code Blues can be predicted in advance, applying medical interventions can prevent some of these deaths. In addition to saving lives, early Code Blue prediction can also provide time for a hospital to plan for interventions and allocate resources accordingly, such that other patients do not suffer because of sudden shifts of resources during Code Blues.

There are some existing early warning systems [4] being used to predict which patients are at risk of entering Code Blue. However, these systems have some shortcomings:

1. They require continuous monitoring of patients by clinical staff. This might be possible for patients in the ICU, but around 50% of Code Blues take place in wards outside the ICU, where monitoring is only intermittent, and the interval between visits could very well be as long as 8 hours [5].

2. They use only limited patient characteristics. With increased use of electronic medical records (EMRs), we believe that using all the existing patient characteristics can result in a more effective early warning system.

3. They classify patients into two classes (positive or negative) without ranking them, making it difficult to prioritize within the group of patients classified as likely to go into Code Blue.

The goal of this work is to flag (and rank) patients who are at risk of cardiac arrest, so that doctors can intervene early before a cardiac arrest (Code Blue event) occurs. The questions we consider are: i) how early can we make a prediction? ii) how precise are our predictions? and iii) how many Code Blues do we catch before they happen?

The sooner we can make a prediction, the sooner the intervention can be administered, and the more time a hospital

has to plan for this and adjust schedules of response teams and availability of equipment. At the same time, our predictions need to balance the false positive rate to avoid unnecessary interventions (and associated opportunity costs) with making sure that we catch and prevent as many future cardiac arrests as possible. These factors can be considered as tradeoffs in our system since we could possibly increase the precision by waiting until a few seconds before a cardiac arrest to predict such events very accurately, but doing so makes the alert mostly useless since it may be unpreventable with just a few seconds of lead time. We model this problem as a prediction problem and use data collected by monitoring different vitals and lab results of patients leading up to the current time to flag patients who are at risk of entering Code Blue in the next several hours. We treat this as a classification problem and use support vector machines and logistic regression. We compare their performance to the early warning method currently used by hospitals (MEWS) and find that our methods significantly outperform the MEWS method and result in earlier, more precise Code Blue predictions. The methods we develop in this paper are focused on a much broader class of patients (ward patients), unlike the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database, which consists of data coming only from ICU patients at a much higher frequency through continuous monitors. This makes our approach especially important for ward patients who are not under constant monitoring but are still at risk of cardiac arrest.

The rest of the paper is organized as follows. We begin by introducing our project partner and data in Section 2. We then explain the model setup and the classification methods used in Section 3. In Section 4 we explain, in detail, our method for feature extraction. We also briefly mention other methods that could potentially improve our method, and are part of future work. In the next two Sections (5 and 6), we explain construction of training and evaluation sets, along with a method for assessing the performance of different models, and for estimating variability in prediction. Finally, we end with our findings in Section 7, and a short discussion about the results and deployment plans in Section 8.

2. DATA AND PROJECT PARTNER

The work described in this paper was done in partnership with NorthShore University HealthSystem. This hospital system contains four individual facilities in the Chicago area. We used patient-level information for about $133,000$ in-hospital patients extracted from electronic medical records (EMR) from 2006 to 2011. Code Blue was called for only 0.5% of the $133,000$ patients in the hospital. Some of the patients had multiple hospital encounters: the data consists of around $232,000$ encounters, and Code Blue alerts were called on only 815 encounters (for patients in the general medical-surgical ward, as well as in the Intensive Care Unit). This extremely skewed class distribution makes the prediction task very difficult.

The data extracted from the EMR system includes information on demographics, past hospitalization history and real-time vitals and laboratory tests for these patients. Demographic features include patient's age, gender, race and ethnicity. We have information on twenty-nine different vitals and lab tests measured at different times since the patient was admitted in the hospital. Some examples of vitals are respiratory rate, systolic and diastolic blood pressures, pulse oximetry and temperature. Lab tests include hemoglobin, platelet count, hematocrit, creatinine and sodium levels. Figure 1 is a summary plot of different vitals and labs in the data.

These measurements are neither taken at equally spaced time points nor at the same frequency for different vitals and lab tests, as demonstrated in Figure 2. We did not perform any sampling and obtained the data exactly as it was recorded in the EMR system. Figure 2 displays the readings of 5 different vitals of a patient who went into Code Blue (at the time indicated by the vertical red line). As we can see in the plot, vitals like heart rate and blood pressure are recorded frequently and periodically. However, vitals like carbon dioxide level are recorded less frequently. Moreover, different patients were admitted in the hospital for different lengths of time. This nature of the data makes the application of typical time series algorithms difficult. We need a good and valid procedure to incorporate temporal property of the data along with information embedded in inconsistencies in data collection. Our method for feature extraction is discussed in Section 4.

We approach this problem as a two stage estimation problem. First stage is to estimate trend and temporal properties of the variables that were recorded at irregular time points in the last 24 hours. Because the length of stay is different for different patients, we use vitals information in the last 24 hours for estimating trends, and use length of stay as a feature. This truncation was necessary to have a uniform comparison across patients. Next, using the estimated trend and temporal properties along with other time constant features, we use standard classification methods like support vector machine and logistic regression, to classify patients who went into code-blue in their hospital stay. We briefly discuss the classification model in the next section.

3. CLASSIFICATION METHODOLOGY

We use classification methods to predict whether Code Blue will occur for a patient in the next h hours, from a given time point, using information leading up to that time. The time point at which prediction is made is denoted as T^P. Time when the event occurs, i.e. h hours after T^P, is called event time and is denoted as T^E. We are interested in predicting the outcome (Y_i) at T^E where:

$$Y_i = \begin{cases} 1 \text{ if patient i goes into Code Blue by } T^E \\ 0 \text{ if patient i does not go into Code Blue by } T^E \end{cases}$$

We create features with medical information up to T^P for both patients who went into Code Blue at T^E (case patients) and those who did not (control patients). Next, given the features at T^P and before, we want to predict how likely a patient is going to go into Code Blue h hours after T^P, such that $T^E - T^P = h$. In our analysis, $h \in \{1, 2, 3, 4\}$, implying that we have four different classification problems. Also note that there is not a well-defined T^E for patients who never go into Code Blue. It ranges from the time a patient was admitted to when the patient was discharged. As the patient never went into Code Blue, the patient can be a negative example in the entire period. We provide an argument on why choice of T^E for control patients matters, and give some intuition on how we went about choosing it for training in Section 5. Figure 3 is a pictorial representation of the problem setup.

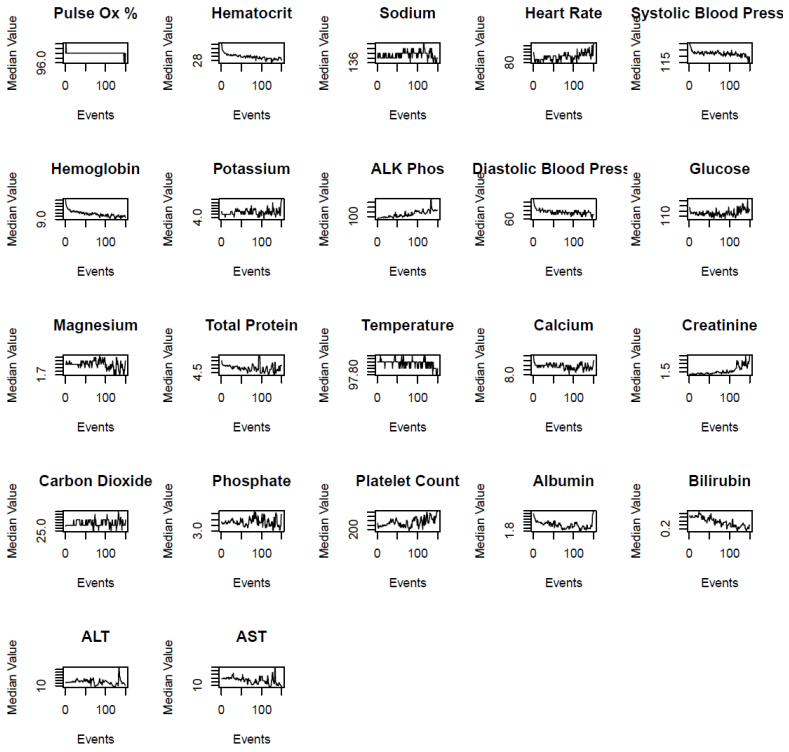

Figure 1: Summary plot of different vitals and labs for patients who go into Code Blue: Along the x-axis are different time points for which data was collected for at least one patient in the last 150 time points. Note that the event times are not regular. Along the y-axis are the median measurements for patients at that time point. This plot shows the variation in population medians of different vitals and lab measurements available for patients right before Code Blue.

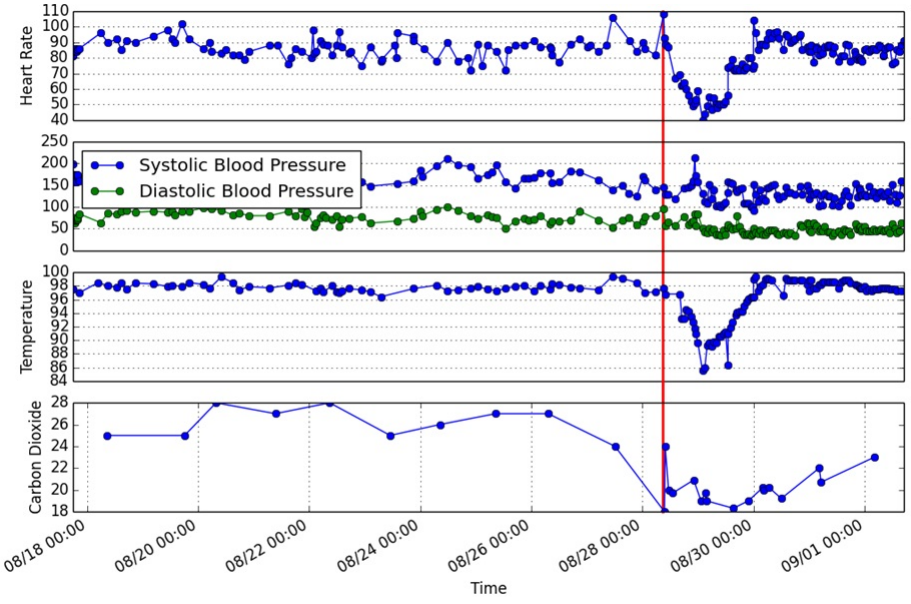

Figure 2: Example of patient-level data for five different vitals. The vertical red line represents the time of Code Blue.

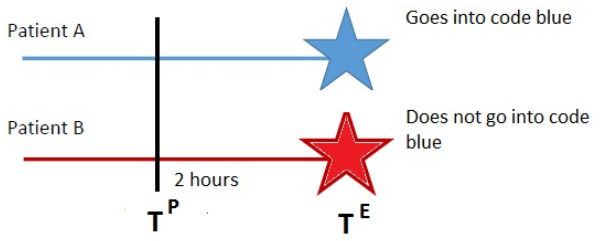

Figure 3: Pictorial representation of the problem setup. We want to classify the events labeled by blue star (=1) and red star (=0), using information before T^P. Time interval between T^P and T^E in this example is 2 hours. We vary prediction time from 1 to 4 hours to see how early we can predict Code Blue.

Support vector machine (SVM) with radial kernel and logistic regression with lasso penalty are used for classification. We train these classifiers on training data, and evaluate their performance by making predictions on held-out (future) test data. SVM with radial basis kernel results in better classification performance but we find that the logistic regression models are more interpretable. While the main goal of this paper is building a good predictive model, it is also of interest to explore how logistic regression perform compared to SVM and what are the important features selected by logistic regression.

Define Y^n as a vector of n outcomes and $X^{(n\times p)}$ as a matrix of p features, for each classification problem. Recall that SVM is formulated as a following optimization problem [11]:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\psi_i \quad \text{subject to } \psi_i \geq 0, y_i(x_i^T\beta+\beta_0) \geq 1-\psi_i \forall i.$$

Here C is the cost parameter that regulates the number of overlaps in training points, and ψ is the proportion by which the prediction is on the wrong side of the margin, also called slack variables. We want to maximize the separation while minimizing the overlap. For non-linear boundary in original space kernelized SVM is used. Radial basis kernel is given by $K(x, x^{'}) = exp(-\gamma||x - x^{'}||^2)$, where γ is a scale parameter. C and γ are tuning parameters chosen by 5-fold cross-validation in training data to find the best model that fits the selection criteria defined below.

Similarly, in logistic regression classification is done by estimating probability of $(Y_i = 1|X_i)$ using the model, $\text{logit}(P(Y_i = 1|X_i)) = X_i^T\beta$. When the dimension of X is large relative to the number of observation, logistic regression with lasso penalty helps to select sparse predictive models [10]. In this framework the coefficients, $\hat{\beta}$s, are estimated by solving following problem:

$$\min_{\beta} -\text{log likelihood } (\beta) + \lambda|\beta|.$$

λ is a tuning parameter of the model which controls the amount of regularization, and hence the sparsity in the estimated coefficients [11]. λ is selected by 5-fold cross-validation on training data to find the best model that fits the selection criteria.

We use F_1 score as a selection criteria. It is a measure of a test's accuracy. F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0. F_1 score is given by:

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Recall (true positive rate) is computed as the proportion of true predicted Code Blue events over all the observed Code Blue, whereas FPR (false positive rate) is the ratio of false predicted Code Blue over total number of events that were observed as non-Code Blue. Since we are interested in predicting Code Blue with high recall and precision, we use this criteria for model selection in cross-validation.

Finally, note that the columns of the features are centered and scaled for these methods.

3.1 Comparison with existing methods used in hospitals

Modified Early Warning Score (MEWS) is a composite index commonly used by hospital staff to determine the severity of a patient's illness. We calculate MEWS by individually scoring the aberration of five vital signs (systolic blood pressure, heart rate, respiratory rate, temperature, and level of consciousness) on a scale of 0 to 3 and summing the resulting scores [4] at the time of prediction. A score of five or more is linked to increased likelihood of Code Blue [9]. We will use 5 as a threshold for predicting whether a patient will go into Code Blue. Because MEWS is frequently used in hospitals as a criteria in activating rapid response teams, it serves as an appropriate basis to compare our results to.

4. FEATURE EXTRACTION

A key aspect in the medical domain (and in any application of data mining) is building effective features for classification. We worked with physicians from NorthShore University HealthSystem as well as researched prior work to determine good features. For Code Blue prediction, it was important to incorporate using irregularly spaced and sparsely collected temporal data. As mentioned earlier, vitals and lab information have atypical properties that hinder in applying regular time series methods for classification:

1. They are not recorded at regular time intervals
2. Some vitals are recorded more regularly than others
3. For a given vital, there is more data at some time intervals compared to others
4. Since patients stay in the hospital for different periods of time, lengths of available data are not uniform across patients

We want to incorporate these properties of the data in our feature set. Along with information like trend and range of the variables at different time intervals, we also want to use information on how often the vital was recorded and when it was missing as features. We believe that they are important to identify the features of the data and of the data collection process that lead to the event.

For the analysis presented in this paper, we start with a simple approach to construct features. We divide the

timeline of the patient's stay in the hospital into different windows to estimate temporal features and trends from the vitals and lab tests that change over time in an encounter [8] [6]. The timeline for temporal variables before prediction time, T^P, was divided into three windows. The first window contains information up to 3 hours before T^P, the second window contains information between 3 to 9 hours before T^P, and the third one contains information between 9 to 18 hours before T^P. We truncated the data at 18 hours for uniformity across patients, since we did not have information beyond 18 hours for some of the patients. We also use total length of stay of a patient as a feature. For each window and each patient, we compute the first and second moments of the empirical distribution of a variable in that window, along with minimum and maximum values, and the frequency of data collection. We estimate the trend in each window by segmented regression [6]. We fit a regression line in each window and record the slope of the line to include increasing or decreasing trend information in feature set.

This process results in 381 features using our data. These include both time constant features and the features we built for time varying variables. Missing values are replaced by the feature's mean for that time window.

We realize that this is still a rather naive approach to estimating trends, but is better than taking the average over all the periods, or considering each measurement of a vital for a patient at different time points as independent (i.i.d) (which they certainly are not). Currently we are fitting separate (discontinuous) least squares in each window that is of uniform size across patients. Ideally, we would like to use all available measurements of a vital for a patient to fit a piecewise linear function that estimates underlying linear trend and use its summary as features for classification [7].

There are more sophisticated methods in the literature (mostly in economics) related to trend filtering in time series or temporal data analysis, that will be a good guidance to further advance this work. We are especially interested in implementing l_1 trend filtering approach proposed by Kim et. al [7] for continuous temporal data. We are currently exploring ways to incorporate this method in our framework.

5. TRAINING

Since our goal is to build a system that will be deployed in the hospital response system, it is critical that the classifier works in a real-time hospital setting. We train our model on historical data, and use it to predict Code Blue on future holdout data. Our goal is to classify patients who will go into Code Blue as early as possible. We train different models for each early warning time threshold: one model that predicts if a patient goes into Code Blue in the next one hour, another in the next two hours, and so on. We choose the best parameters through 5-fold cross-validation of each model on training data.

One of the important considerations in our classifier is to choose the number of negative examples that should be considered to learn the differences in the two classes (Code Blue and non-Code Blue). Note that this is a highly unbalanced classification problem, where Code Blue is a rare event. We use a nested case-control approach to match control patients to every case patient, so that a classifier can learn the differences in training data easily. We match Code Blue patients to non-Code Blue patients with similar demographic features only, such that there remain differences in their physiological and hospitalization histories. We create a training set that is less skewed by sampling the data, to account for the unbalance of the two class labels. We vary the ratio of cases to controls in our training set as 1:1, 1:4 and 1:10, and we decide on the appropriate ratio based on the prediction results from the training data. It is important to note that we do not change the distribution of the test set. We only vary the number of controls sampled to construct the training set for our classifiers, but we use the natural distribution of the data for the test set. This is done to make sure the models we build are applicable when put in production and the results are still consistent with the initial experiments.

Further, we need to select time of comparable event (T^E) for control patients, so that we can build up the negative examples. For patients who go into Code Blue, time when they go into Code Blue defines the positive event. However, for patients who never go into Code Blue, we have the time from when they were admitted to the time when they were discharged as a choice for the negative event. Most of the patients who get discharged are healthier at the time of discharge than they were during (or at the beginning of) their stay. This makes using time of discharge as a comparable event a bad choice, as the classification problem will become artificially easy. Using time close to admission will not give us enough information to build an accurate model. In this analysis, we vary the time of event for control patients, as the 25th, 50th and 75th percentile of their stay time in the hospital, and select the value that performs best in the training data.

We experiment with using support vector machine (SVM) with radial kernel and sparse logistic regression with lasso penalty as classifiers [11]. γ and C for SVM and λ for logistic regression are selected by cross-validation with maximizing F_1-score as a criteria. Once we select optimal parameters, we refit on the entire training set using these tuning parameters for each method, and use the estimates for prediction on test data.

6. EVALUATION AND ESTIMATION OF STANDARD ERRORS

For every hour in the evaluation set, we take the features of all the patients in the hospital present at that hour. Using these features, we predict which patients would go into Code Blue 1, 2, 3 and 4 hours from that time point. This evaluation is used to compare classifiers and compute the best parameters for those classifiers, using recall and false positive rate (FPR) as metrics. Recall is computed as the proportion of true predicted Code Blue events over all the observed Code Blue, whereas FPR is the ratio of false predicted Code Blue over total number of non-Code Blue events. Also, receiver operation characteristics (ROC) curve for aggregate monthly predictions along with area under the curve (AUC) are compared among classifiers to evaluate their performance.

Our first training set consists of data before January 1, 2011. We use events in the month following the training period as test data. We then sequentially add additional months to the existing training data, and hence create 7 new training data. Estimates from each training set is evaluated on the test data in the following month. Following this procedure we have 8 training and test data. We compute mean

recall and FPR, along with the standard error of the mean, over the 8 test data. We also compute MEWS scores for patients in the test data. Patients with MEWS scores greater than or equal to 5 are predicted to go into Code Blue in the next 4 hours. Note that, the first training data is contained in the next training set and so on. This is a time dependent data, where the Code Blues in the past might affect how hospital deals with Code Blues in the future. Hence, we believe that the hospital system learns from the past, so we use the past data to predict the future events. At the same time, multiple training and testing datasets give us a valid way to estimate standard errors of the mean recall and false positive rates for different models. Also, receiver operation characteristics (ROC) curve for aggregate monthly predictions along with area under the curve (AUC) are compared among classifiers to evaluate their performance.

7. RESULTS

Based on mean recall and false positive rate from predictions on multiple test data, we find that using 1:4 case to control (positive to negative class) ratio and 75% percentile time point as a comparative event for the control group in training gives the best results in terms of AUC. The rest of the results presented in this section are based on using these parameters.

We were able to predict Code Blue with around 80% recall and 20% false positive rate at 4 hours ahead of the event. We find that SVM performs better than the current method used by medical practitioners (MEWS) in terms of mean recall and FPR, as shown in Figure 4, and in terms of ROC and AUC, as shown in Figure 5. We also ran sparse logistic regression on the similar framework. Prediction from SVM has significantly lower false positive rate compared to logistic regression. Even though predictions from SVM have higher mean recall, they are not significantly higher for all time points. Both, SVM and logistic regression performed better than MEWS in terms of recall. False positive rate estimated from MEWS has bigger standard errors, and thus we could not be certain about the performance of our method compared to MEWS. As we can observe from the results, the benefit of using our machine learning methods increases as we make our predictions earlier. MEWS gets much worse (especially in terms of recall) compared to SVM as we make our predictions earlier in time. This is critical since our goal is to make these predictions as early as possible to enable the hospital to intervene early, as well as modify the schedules for physicians and nurses to optimize overall quality of care for all patients.

One of the advantages of logistic regression is its popularity among medical practitioners because of its interpretability in terms of understanding important features. From our analysis, we see that the mean, the variability, and the trend (rate of acceleration/deceleration) of features like heart rate, respiratory rate, diastolic blood pressure, and pulse oxygen levels are important features for predictions at all hours. Moreover, the frequency at which lab measurements are taken is also a good predictive feature, and its importance increases as we make earlier predictions. Figure 6 shows the top features selected in the 1 hour ahead prediction. One interesting finding is that as we try to predict further out into the future, more features become relevant. For example, 72 features were selected as important features in the 1 hour ahead as compared to 100 features for the 2 hours ahead prediction. We think that this is because prediction gets more difficult as we try to predict Code Blue earlier, and the model needs (and uses) more features to perform effectively. However, we do not have any measure of uncertainty of these estimates, and while these features are helpful to understand their effect on Code Blue, it is difficult to make any inferences. The important features selected by our model are shown in Figure 6, where we also provide the coefficients of features in the sparse logistic regression. The coefficients represent the log odds ratio of the feature in predicting Code Blues.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we described a system we are building to create an early warning prediction system for in-hospital cardiac arrests that surpasses the existing methods used by hospitals and medical practitioners. This gives hospital staff the ability to intervene before the patient becomes critical, thus averting some potential Code Blue occurrences, and ultimately saving more lives. This also reduces the opportunity cost of the medical personnel having to drop their current hospital activities to rush to attend to Code Blue patients, and facilitates more efficient and effective planning and scheduling. We found that this prediction can be made as early as 4 hours before the Code Blue event with high recall and relatively low false positive rate. Most importantly, we verified the hypothesis that patients show clinical signs of deterioration before going into cardiac arrest, and vital signs and lab measurements can be used to determine these signals.

There are certainly approaches that would improve the performance of these models by adding additional features that we did not have access to, such as using diagnosis information from patient records. Also, the extreme imbalance in class labels in the data had caused some difficulties in creating test and training sets. A good future direction would be to combine data from different hospital systems, so the algorithm can be trained on more positive examples, which should improve predictions further.

While we have significantly higher recall compared to the currently used score (MEWS), we are only slightly better than MEWS in mean FPR. A future step would be to work with medical practitioners and understand this trade-off in deployed systems.

Lastly, as mentioned in Section 4, there is a lot of potential for improving estimation with trends in time-varying features. Exploring more sophisticated and statistically sound time-series methods is a priority for future work.

9. DEPLOYMENT PLANS

The work described here was done in partnership with NorthShore University HealthSystem. We collaborated with the team in the hospital and presented these methods and results to a set of doctors and researchers. Based on our results that outperform the currently used MEWS score, we are now in discussions with NorthShore Hospital as well as with other hospitals and analytical providers to determine the best way to deploy such a system. One approach we are considering is converting the classifier into a set of database queries that runs on top of the EMR system and feeds into an alerting system. This alerting system would raise an alert when a potential Code Blue is about to happen in the next

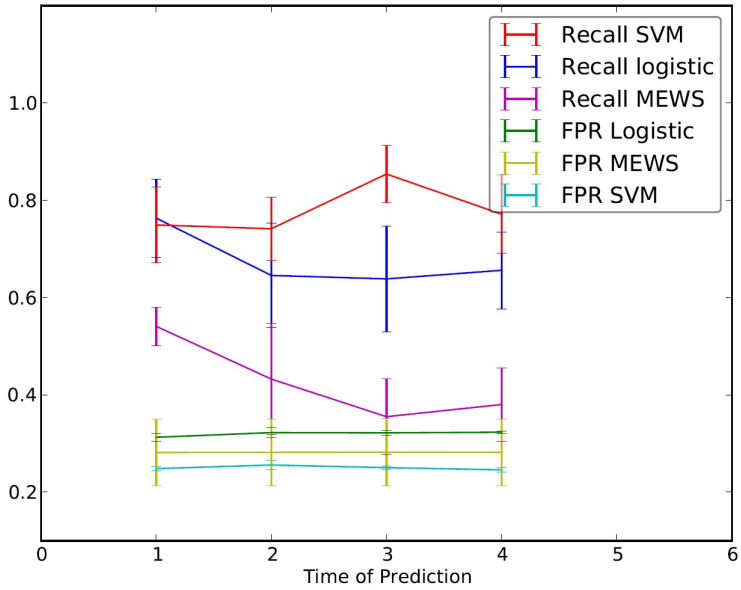

Figure 4: Comparison of the performance of SVM, Logistic Regression and MEWS. Along the x-axis is time of prediction before Code Blue. For example, 4 on the x-axis means the prediction is made 4 hours before the event. Along the y-axis is recall or false positive rate. Different colors are for different classification methods.

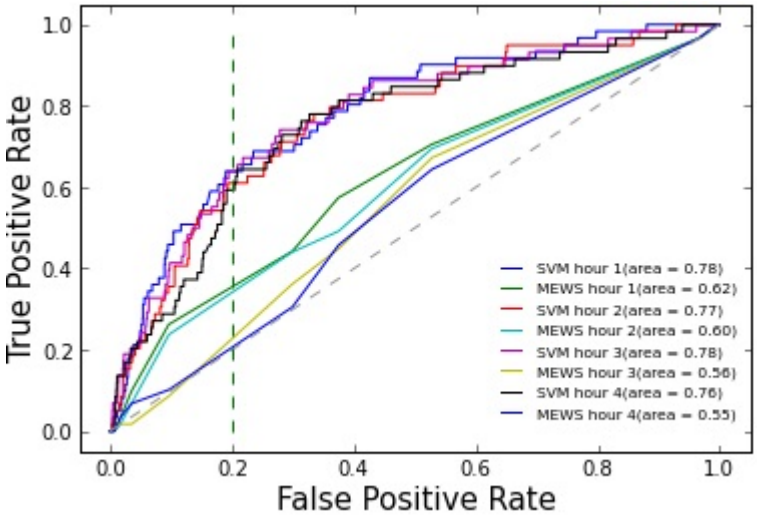

Figure 5: ROC curves for comparing the performance of SVM and MEWS. Area under the curve for each method, along with the number of hours ahead when making the prediction, is mentioned in the legend.

Number of Previous Code Blues	0.23
Number of Previous Discharges to Home	0.16
Emergency Admission	0.15
Number of Previous Encounters	0.13
Minimum Potassium	0.24
Platelet Count Window 0	0.22
GFR Count	0.12
Glucose Count	0.19
Glucose Count Window 2	0.11
Hematocrit Count Window 1	0.16

Average Respiratory Rate	0.20
Average Heart Rate Window 1	0.11
Minimum Heart Rate	0.14
SD Respiratory Rate Window 1	0.29
SD Heart Rate Window 0	0.13
SD Heart Rate Window 2	0.21
SD Diastolic Blood Pressure Window 2	0.16

Figure 6: The top features selected by our sparse logistic regression for the 1 hour ahead prediction.

n hours. In addition, the classifier would also provide input to the hospital scheduling system that will allow administrators to optimize the overall quality of care they provide to patients.

10. ACKNOWLEDGEMENTS

This work was supported by the 'Eric and Wendy Schmidt Data Science for Social Good' Summer Fellowship at the University of Chicago in 2013. We also thank Dr. Jonathan Silverstein, Vice President and Head of the Center for Clinical and Research Informatics at NorthShore University Health Systems, for his valuable inputs, access to data and supporting our efforts to deploy the methods into their hospitals.The data from this study were obtained as part of a collaboration between the University of Chicago and NorthShore University Health Systems, which was funded by an institutional Clinical and Translational Science Award grant (UL1 RR024999; PI: Dr. Julian Solway). We are grateful to Drs. Dana Edelson, Matthew Churpek, Ari Robicek and Christopher Winslow for study conception and/or data acquisition. In addition, Poome Chamnankit, Kelly Bhatia, Audrey Seitman and Justin Lakemen extracted the data.

References

[1] C. A. Alvarez, C. A. Clark, S. Zhang, E. A. Halm, J. J. Shannon, C. E. Girod, L. Cooper, and R. Amarasingham. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Medical Informatics and Decision Making*, 13:28, 2013.

[2] M. Buist, E. Jarmolowski, P. Burton, S. Bernard, B. Waxman, and J. Anderson. Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. a pilot study in a tertiary-care hospital. *Med Journal*, 171, 1999.

[3] P. Chan, A. Kalid, L. Longmore, R. Berg, M. Kosiborod, and J. Spertus. Hospital-wide code rates and mortality before and after implementation of a rapid response team. *JAMA*, 300, 2008.

[4] M. M. Churpek, T. C. Yuen, M. T. Huber, S. Y. Park, J. B. Hall, and D. P. Edelson. Predicting cardiac arrest on the wards, a nested case-control study. *American College of Chest Physicians*, 11-1301, 2012.

[5] D. A. Jones, M. A. DeVita, and R. Bellomo. Rapid-response teams. *New England Journal of Medicine*, 365(2), 2011.

[6] E. Keogh, S. Chu, D. Hart, and M. Pazzani. Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases*, 2004.

[7] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. l1 trend filtering, 2009.

[8] C. Sian and L. David. Effective probability forecasting for time series data using standard machine learning techniques. *Proceeding of the third international conference on advances in pattern recognition*, August 2005.

[9] C. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, 2001.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[11] H. Trevor, T. Robert, and F. Jerome. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.