Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records

Karla Caballero, PhD Candidate\textsuperscript{1} and Ram Akella, PhD\textsuperscript{1,2}

\textsuperscript{1}University of California Santa Cruz, 1156 High Street Santa Cruz CA, USA
\textsuperscript{2}University of California Berkeley, 94720 Berkeley CA, USA

Abstract

In this paper, we propose a framework to dynamically estimate the probability that a patient is readmitted after he is discharged from the ICU and transferred to a lower level care. We model this probability as a latent state which evolves over time using Dynamical Linear Models (DLM). We use as an input a combination of numerical and text features obtained from the patient Electronic Medical Records (EMRs). We process the text from the EMRs to capture different diseases, symptoms and treatments by means of noun phrases and ontologies. We also capture the global context of each text entry using Statistical Topic Models. We fill out the missing values using a Expectation Maximization based method (EM). Experimental results show that our method outperforms other methods in the literature terms of AUC, sensitivity and specificity. In addition, we show that the combination of different features (numerical and text) increases the prediction performance of the proposed approach.

Introduction

Currently, the accurate and opportune prediction of patient readmission to the ICU shortly after he is transferred to a lower level care is of great interest to health providers. Tools that estimate the probability of patient relapse and readmission to the ICU aid physicians and health care providers to determine the possible resources that should be allocated to the patient and to discover the possible causes of relapse that could lead to the patient to be readmitted to the ICU. The timely and accurate estimation of patient’s probability of readmission allows us to successfully trigger a medical alarm before the patient is discharged from the ICU.

This probability estimation also permits the early identification of patients with elevated risk of readmission. As a result health care providers can differentiate those patients from the ones who are stable and less likely to return in order to assign medical resources more effectively.

Most of the existent methods in the literature\textsuperscript{1,2} rely on the use of static classifiers that do not take into account the evolution of the patient nor the dynamic nature of the patient’s features. To overcome these challenges, we propose a dynamic method based on Bayesian Time Series and Dynamic Linear Models (DLM) to estimate the probability of readmission before the patient is discharged to indicate the existence of a possible medical alarm. Our contribution is summarized as follows: We model the probability of patient readmission as an aggregated latent state which is updated each time new features are observed (lab results, vital signals, etc.). Our model is fed with heterogeneous data obtained from his Electronic Medical Records (EMRs) which consists of text and numerical data with both discrete and continuous variables. We incorporate the text information into the model by developing a method which converts the unstructured text information into discriminative features that are later incorporated into the model. Finally, we address the missing values problem by estimating those values using a Regularized Expectation Maximization (EM) based method.

In this context, we find that the dynamic estimation of the probability of patient readmission to the ICU provides a more accurate prediction when compared with other methods in the literature and other static prediction methods.
Background

Prevailing medical practice relies on frameworks such as the Apache III\(^1\), and SAPS II\(^4\) scores. Both methods, widely used to predict patient mortality in the Intensive Care Unit (ICU) are used as proxy to estimate the likelihood that a patient is not ready to be transferred to a lower level care and there is a high probability of being readmitted if he is discharged from the ICU. These methods incorporate temporal information in a limited way by only choosing the worst-case scenario values during the first 24-hour window that a patient is inside the ICU. As a result, they often overestimate the probability of mortality (not readiness to be transferred from the ICU). Moreover, these scores are only estimated once during the entire stay in the ICU, which may not indicate whether the patient would recover and be successfully transferred from the ICU in the future.

Data Mining has been previously used to address the problem of estimating the likelihood that a patient is readmitted to the ICU\(^1,2\). Other prediction methods such as Batal et al.\(^5\) include the dynamic information by collapsing the time series of features, such as blood pressure and heart rate, into static features that are later used in a classification framework. However, this model does not take into account the evolution of the patient in time Most of the existing solutions rely on training a static classifier with a patient’s observed feature vector. These features are mainly static, such as lab reports produced at a given point of time or the estimated APACHE score at the discharge time.

Furthermore, most of the methods mentioned above assume the availability of all the features at the prediction time. This assumption may not be valid in a real scenario where the data is often incomplete and segmented. Health care data suffers from a large volume of missing data due to the fact that not all the features are collected (lab results, vital signal, etc) for all the patients at all time. One of the most common methods to fill out these missing values is to perform mean imputation. However, this practice has been shown to introduce more noise into the model rather than reduce it\(^6\). To tackle this problem, previous approaches segment the patient features according to their age group and then calculate the average value for each segment\(^7,8\). Other methods handle missing values by fitting a distribution for each feature with the observed data and sample from the estimated distribution when the value is missing\(^9\). Similarly, the use of Multiple Imputation to predict the missing values has been proposed previously. Here regression techniques with the other observed features as covariates are deployed\(^6,7\). Overall, these methods do not take into account the temporal aspect of the missing data where current features values are highly dependent on previous values.

Most of the existing prediction models do not use text from the Electronic Medical Record (EMRs) due to its complexity. However, text data contains key information that is potentially useful to better predict the likelihood that a patient is readmitted to the ICU. Examples of text include lab reports, admission, doctors and nurse notes. Ghassemi et. al\(^10\) combine static numerical features such as SAPS II score with topic modeling features from the text of the EMRs to estimate the probability that a patient die after 30 days of being discharged using Support Vector Machines (SVM).

Our proposed approach combines text and numerical information in a dynamic setting that allows us to predict the patient readmission before the patient is transferred from the ICU to other hospital areas. In addition, our proposed approach takes into account the evolution of the patient health state and temporal aspect of the patient features to predict how likely he will relapse and will be readmitted to the ICU.

Methods

In this section, we describe the method to construct the probability of readmission as a latent state, the methodology we use to extract numerical features and to process the text information to extract the discriminative features. In addition, we outline the framework we use to handle missing values. Finally, we describe the experimental settings and the methodology to validate the proposed approach.

Definition of Probability of Readmission as a Latent State

We define the that a patient \(i\) would be readmitted to the ICU in the next 30 days \(Y_{t,i}\) or not \(Y_{t,i} = -1\) if it is discharged from the ICU at time at time \(t\) as a binary variable. \(Y_{t,i}\) is 1 with a probability of \(\pi_{t,i}\), which represents the probability that a this patient is readmitted if he is transferred from the ICU at time \(t\) (probability of readmission). This probability is a function of a latent state \(\theta_{t,i} = [\xi_{t,i}, \tilde{\xi}_{t,i}].\) This latent state is formed by the estimation of the log-odds of the probability of readmission in previous steps \(\xi_{t-1,i}\) and the patient features observed at time \(t\), \(\tilde{\xi}_{t,i}\). The value of \(\tilde{\theta}_{t,i}\) is obtained by combining a set of observed features \(X_{t,i}\) obtained from the EMR at time \(t\) and the value
of this combination at previous steps, $\tilde{\theta}_{t-1,i}$.

In this framework, we are able to include both the patient’s features and his health context obtained from previous time steps. This is not accounted for in the static classification frameworks. Our proposed model is a special case of the Generalized Dynamic Linear Models (GDLM)\textsuperscript{11}. Here, we employ the logistic transformation to accommodate our specific context. This leads to the following expressions:

\begin{align*}
Y_{t,i} & \sim \text{Bernoulli}(\pi_{t,i}), \quad \pi_{t,i} = \frac{e^{\xi_{t,i}}}{1 + e^{\xi_{t,i}}}; \\
\xi_{t,i} & = \xi_{t-1,i} + \tilde{\theta}_{t,i} + w_{t,i}; \\
\tilde{\theta}_{t,i} & = \lambda\tilde{\theta}_{t-1,i} + \beta X_{t,i} + \phi_{t,i}; \\
\phi_{t,i} & \sim N(0,W_\phi), \\
\xi_{t,i} & \sim N(0,W_\xi).
\end{align*}

Here $\lambda$ is a decay factor that determines the contribution of previous feature values in the current value of $\tilde{\theta}_{t,i}$. The vector $X_{t,i}$ is constructed from the patient’s observed lab test results, vital signs readings, text notes features, and demographics (features). In this model we assume that most of the values of $X_{t,i}$ are observed. In later subsections of this paper, we explain how we model and estimate the missing values of the feature vector. The vector $\beta$ represents the regression coefficients we use to combine the observed features. The value of $\theta_{t,i}$ can take both positive and negative values. Thus, we are able to increase or decrease the probability of readmission using the observed features $X_{t,i}$. $W_\xi$ and $W_\phi$ are the evolution variances of $\xi$ and $\theta$ respectively.

The value of $\xi_t$ reflects the log-odds effect on the probability of readmission $\pi_{t,i}$ by previous observed features contained in the state $\theta_{t,i}$. To illustrate this effect, we calculate the impact of the user’s features $X_{t,i}$ observed at time $t$ and then aggregate them into the state $\theta_{t,i}$ after $k$ steps assuming no other values of $[X_{t+1,i}, \ldots, X_{t+k,i}]$ are observed. This impact is determined by the following forecast function:

$$
\xi_{t+k} = \sum_{r=0}^{k} \lambda^r \tilde{\theta}_r = \tilde{\theta}_t(1 - \lambda^{k+1})/(1 - \lambda)
$$

As illustrated by the previous equation, the proposed model incorporates knowledge from prior measurements into the current state estimation. This effect representation allows us to predict patient probability of readmission even when no measurements are available at a given time $t+k$. In addition, the effect does not decrease over time, as opposed to $\theta_t$ (observed features). Each time there are new observations available, the value of the effect $\xi_{t,i}$ is updated using equation 4.

**Model Fitting**

Figure 1 shows the graphical model of this framework. The colored circles represent the variables that are observed in the model. The non-colored circles are the latent variables and model parameters that need to be inferred. The learning across multiple users is reflected through the estimated parameters $\Phi$ defined as: $\Phi = \lambda, W_\phi, W_\xi, \beta$. This representation is flexible enough to expand the model and to incorporate different weighting vectors $\beta$ for different patient groups with a particular disease or age range.

We fit the model using Dynamic Linear Models with Logistic Transformation. This model incorporates the user features into an aggregated patient state that evolves over time, in contrast to static classification models. In addition, we train the model using the entire patient’s stay path as opposed to individual time steps. By performing this, we take into account the uncertainty about the future in the estimation of the probability of readmission.

Our proposed approach allows us to predict future values of the state as more readings become available. Consequently, we are able to dynamically estimate the current patient probability of readmission and predict its evolution using the predictive forecast function of the latent state. Figure 2 describes the fitting steps for the proposed model.

In the following subsections, we describe the process to extract both numerical and text features, as well as the method to fill out missing values. Once we have the time series features for all the patients in the training set, we give an initial value to the model parameters $\Phi$.

We train the model using an iterative method based on Expectation Maximization (EM). This method consists of 2 steps: E and M steps. In the E step, we estimate the latent state of the patient $i$, $\theta_{t,i}$ using the Forward Filtering
Backward Smoothing method (FFBS)\(^1\). In this method, we estimate first the latent state \(\theta_{t,i}\) using the values of the observed features \(X_{t,i}\) and the state value of the previous time step \(\theta_{t-1,i}\) (Forward Filtering). Once we have estimated the entire path, we correct the estimated latent state backwards using the estimated state values of future time steps \(\theta_{t+1,i}\) (Backward Smoothing). By combining the Forward Filtering (FF) and the Backward Smoothing (BS), we guarantee the construction of a fully dynamic model with feedback where previous values of the path affect the current latent state while accounting for the future uncertainty. One variant of this dynamic model is to train the model with no feedback about the future (open loop feedback). To achieve this, we fit the model using the latent states obtained with the Forward Filtering (FF) step only.

The M-step consists of estimating the values of the parameters \(\Phi\) that optimize the latent patient paths (probability of readmission path) previously estimated in the E-step. Then, we repeat the E and the M steps until convergence.

We predict the probability of readmission in the test partition assuming that we do not know the final outcome of the patient. Therefore, we estimate the latent state \(\theta_{t,i}\) using the estimated parameters and previous values of the latent state \(\theta_{t-1,i}\) using Forward Filtering.

**Numerical Features Extraction**

We extract the numerical features using the events described in the EMR. A patient may be subject to different procedures and events during his stay in the ICU based on his condition. The events that we incorporate into the model are selected by means of the \(\chi^2\) test. We estimate then this score for all the events that appear in the corpus and retain those with higher score (the most discriminative ones). We extract 30 features such as blood pressure level, lab procedures, pain level, and heart rate.
Table 1: Feature Extraction Process

<table>
<thead>
<tr>
<th>Numerical Features</th>
<th>Term-Based Features</th>
<th>Topic-Based Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Transform bimodal features using Apache Score III weights</td>
<td>• Construct term frequency matrix</td>
<td>• Fit topic Model</td>
</tr>
<tr>
<td>• Perform $\chi^2$ test on all the observed features</td>
<td>• Estimate the $\chi^2$ score for all terms and retain those with highest score</td>
<td>• Estimate topic Mixture for every text entry</td>
</tr>
<tr>
<td>• Select the features with higher separation score</td>
<td>• Classify each text entry as should continue in the ICU (1) or ready to be discharged (-1) using the selected term features</td>
<td>• Determine presence or absence of each topic</td>
</tr>
<tr>
<td></td>
<td>• Classify each text entry as should continue in the ICU (1) or ready to be discharged (-1) using topic features</td>
<td></td>
</tr>
</tbody>
</table>

We observe that the selected features are a combination of those used by the APACHE III and SAPS II scores\(^3,4\). We find that 80% of the Apache III score features and 90% of the SAPS II features are included in our selected features. This selection shows consistency between our proposed approach and these widely used methods of the literature.

We identify that some of the selected features are considered to have a bimodal distribution. For instance it is equally dangerous to have a really low blood pressure as to have it to be really high. To integrate this knowledge, we assign a weight for each possible range of the event. Those weights are obtained from the those used by the Apache Score III\(^3\).

We then divide the selected user features into two groups: quasi-static and dynamic. Some of the labs and procedures do not need to be performed at each time step. Therefore, we consider these features as quasi-static (they are updated if there is a new reading). Features such as blood pressure, pain level and heart rate are considered to be dynamic. This division impacts how missing values are treated in these features. The quasi-static features are updated when a new value is observed. Meanwhile, the dynamic features will be filled in using the method to estimate missing values described later in the paper. In addition to these features, we include two demographic features about the patient: gender and age which we treat as static variables. Both features provide us the initial conditions to set the initial probability of readmission. Table 1 shows a summary of the feature extraction process.

**Text Feature Extraction**

Standard approaches used as proxy to estimate the probability of readmission, such as Apache III and SAPS II scores, do not incorporate text information. One of the main challenges researchers face is to incorporate this type of data effectively and to create discriminative features.

To achieve this, we extract text features to improve the health state prediction, consequently the estimation of the probability of readmission to the ICU. The text entries found in an EMR mainly consist of nurse’s entries, procedures reports, admission and discharge information, among others. Each text entry has an assigned timestamp. Thus we are able to construct a time series for each of the text features we extract. In this subsection, we describe the steps we follow to process the text and extract different features that are later integrated into the model. Figure 3 depicts the text based feature extraction process.

**Noun Phrases Extraction**

Given the nature of health care text data, we need to extract meaningful phrases and concepts to successfully represent diseases and treatments (features) which help us to improve our statistical estimates. To achieve this task, we extract noun phrases relevant to the medical domain by annotating a set of discharge summaries using the Clinical Text Analysis and Knowledge Extraction System (cTAKES)\(^13\) and Metamap\(^14\). With these tools we extract clinical named entities (concepts) such as drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures.

Discharge Summaries provide us a rich data set to explore and to extract noun phrases. These documents often aggregate the patient’s medical history. This includes all the patient’s information collected during his stay in the ICU, treatments and care information that the patient should follow after he is discharged from the ICU.

After extracting all the entities, we select the phrases which describe a disease, a procedure or a medication using the medical ontologies provided by SNOMED\(^15\) (matched concepts). In addition, we also detect which set of noun
phrases corresponds to stop words (i.e. patient name, doctor name). By means of tf-idf term selection, we select the most important noun phrases and remove those with low score.

Once the phrase selection is completed, we perform standard stop words removal and stemming before indexing the documents. Then, we extract two types of features: term and topic based features which we describe below.

**Term-Based Features**

We incorporate into the model a term-based feature using the obtained noun phrases and word terms from the EMR text entry. This feature is based on the classification of the text entries: $-1$ if the patient is ready to be discharged and $1$ if not. We use the Naive Bayes classifier, which has been shown to provide good predictive performance and it is computationally feasible for the unbalanced classes problem\(^{16}\).

In order to make this classification feasible, we reduce the large vocabulary size of the corpus by extracting the most discriminative terms by means of $\chi^2$ test of the probability of readmission\(^{17}\). We performed this test on the whole corpus. Our goal is to obtain a global estimate from the whole corpus in order to reduce the bias resulting from the term selection. We keep 4000 terms after this step.

**Topic-Based Features**

The second set of text features is based on statistical topic modeling. These models allow us to reduce the dimensionality of the term space to a smaller feature space of latent ”topics”. In addition, we are able to model topics for unseen documents without training the model again, as the method is generative.

In this context, each document is represented as a mixture of topics with a certain probability. Similarly, each topic is represented as a mixture of words. Our hypothesis is that topics capture the global context of the document while this cannot be achieved by selecting text terms alone. By capturing this context, we are able to improve the performance of the prediction of patient readmission.

For this application, we fit a GD-LDA model to extract the topics from the corpus set using 75 topics. The corpus consists in all the processed text entry notes (noun phrases + terms) of all the patients. GDLDA, which is a generalization of LDA\(^{18}\), allows us to model correlations between topics as opposed to LDA. In addition, this method is fitted in an unsupervised form, and it is computationally efficient, which permits us to train a large number of documents in a single batch contrary to other statistical topic models that model correlations such as Correlated Topic Models (CTM)\(^{19}\).

We then remove the background topics which we define as word mixtures with a high percentage of common words (more than 90% of the terms inside the topic). We define as common words those that do not have healthcare related information by comparing them with the ontologies from the UMLS using MetaMap\(^{14}\). These ontologies provide
information about healthcare treatments, drugs and diseases. We keep 65 topics after this step.

After removing the background topics, we select the 10 most discriminative topics by means of the $\chi^2$ test and include them in the dynamic model as features. Our hypothesis is that patients with certain topics in their EMR are more susceptible to be readmitted to the ICU than those whose EMR does not have them. In order to make the documents comparable, we use the values of $\{1, 0\}$ to show the presence or absence of a topic in the document instead of the probability of the topic in the document (two patients with similar medical history can have the same topics in their EMRs, but in different proportions). Thus, we indicate that a topic is present in a document if it accounts for more than 5% of the total topic mixture inside the document. In addition to the most discriminative topics, we include the classification output of the text entry (patient ready to be discharged or not) using the document topic mixture as features. Here we use Naive Bayes classifier.

**Missing Features Estimation**

The proposed framework mentioned above assumes that most of the patient’s features are observed at each point in time. When feature values are not observed, we indicate they are missing and then we impute their value.

For the current application, the patient’s features have an implicit temporal aspect. The value of features in time $t$ are highly dependent on their value in previous time steps $1..t−1$. Then, standard imputation methods based on the mean value could lead to estimation errors. To overcome this challenge, we impute the missing values by means of a Regularized Expectation Maximization method. This iterative method uses the observed features at a particular time to impute the missing ones using an initial value of the parameters (E-step). Then we choose parameters that minimize the error rate between the observed values and the imputed ones (M-step).

The E-step (expectation) consists of estimating the missing features $\hat{x}_m$ using the values of the available ones $x_a$ inside the record using the following equation:

$$\hat{x}_m = \hat{\mu}_m + (x_a - \mu_a)\gamma, \quad \gamma \sim N(0, \Sigma)$$  \hspace{1cm} (5)

where $\hat{\mu}_m$ is the mean estimate of the missing features of a record and $\mu_a$ is the mean estimate for the available features in a given record. We define as a record all the observed features of a patient at a given point in time $X_{t,j}$. The value of $\gamma$ is the vector of regression coefficients for the available features. The hypothesis behind this step is to represent the missing record values as a combination of the available values for a given record and an estimated mean of the
missing values of all the records.

After the missing values have been imputed, we perform the M-step. In this step, we estimate the mean and variance of the observed features and coefficient weights of how these are combined to impute a missing value. We repeat both steps until a certain error rate is achieved (around 5%). More details of the algorithm can be found in the work of Schneider.\textsuperscript{20}

**Experimental Settings and Validation**

We test our approach by estimating the probability that a patient discharged from the ICU and transferred to other areas of the hospital would be readmitted in the near future. This prediction is critically important in the accurate prognosis of patient health state (we want to avoid health complications in the patient by discharging him before time). In addition, this prediction is very important to assign medical resources effectively.

For the situation which we are modeling, we predict the probability that the patient would be readmitted if he is discharged at time \( t \) using the information available in his EMR before he is discharged. The EMRs are obtained from the MIMIC II data set.\textsuperscript{21} This dataset contains text and numerical information that describes procedures, medications and vital signs readings from a given patient during his stay in the ICU. MIMIC II is composed of medical records from over 30,000 patients admitted to the ICU during a 7 year window of hospitals from Boston Area.

We validate our method using the EMRs of 15,000 patients selected randomly. In order to compare our approach with other methods from the literature, we only study the adult patients (over 18 years of age) without excluding patients due to a specific illness; this data consists of 11,648 people. We report our results using five-fold cross validation\textsuperscript{1}, taking 80% of the patients as training set and the remaining 20% as test set. We construct the time series of each patient by aggregating the patient information every 3 hours.

We report the probability of patient readmission after \( t = 24 \) and 48 hours and at the time of discharge. Our goal is to test the prediction capability of the proposed approach at different time steps. We compare our method with related methods of the literature used as proxy to estimate the probability of readmission such as Apache Score\textsuperscript{3} and SAPS II\textsuperscript{4}. In addition to these methods, we compare the proposed approach with static classification methods such as Naive Bayes and Random Forests with 50 trees with a sliding window for \( t = 24, 48 \) and before patient discharge as in our proposed method.

**Results**

When analyzing the time series of the patients, we detect that the 57% of the patients stayed 24 hours or less. Thus common practices used as proxy to detect the probability of patient readmission such as Apache Score and SAPS II are not applicable for those patients. We also detect that after dividing the features in quasi-static and dynamic features 34% of the features are not observed. Therefore, the method to fill out missing values would have high impact in the estimation of the results.

Figure 4 shows the obtained topics after constructing noun phrases and removing background topics. Note that by performing these steps we are able to obtain cleaner topics, each one related to a particular disease describing symptoms and body parts than those provided by Ghassemi et al.\textsuperscript{10} In addition to the quality of the topics, we also evaluate the quality of the topic-based classification output using the document topic mixture as features. Here we observe that 72% of the notes from the discharged patients that were readmitted were classified correctly and 52% of the notes from the patients that were not readmitted to the ICU were classified correctly using topic-based features. In addition, we evaluate the performance of the classification using term-based features. Here we observe that 53% of the notes from readmitted patients were predicted correctly and 10% of the notes from patients non readmitted were classified correctly. When comparing both features, we observe that topic-based classification is more accurate than term-based classification. In this context, we corroborate our hypothesis that topics capture the global context of the of the text notes and improve the classification performance when compared with term-based classification.

Table 2 shows the AUC, sensitivity and specificity for the proposed method and other methods proposed in the literature. Here we observe that our proposed framework outperforms the other reported methods in the three reported

\textsuperscript{1}K-fold cross validation provides a more robust prediction evaluation than leave-one-patient-out-method because K-fold cross validation uses less training data

1838
In this paper we have proposed a dynamic model that combines heterogeneous data form the patient Electronic Medical Records to predict the patient readmission to the ICU. The accurate prediction of patient readmission would lead to an good allocation of resources that potentially could reduce cost in the transfer and patient care.

The use of an aggregated patient state which combines current features with previously observed ones allows us to predict the probability of readmission even if not new features are observed. The results of the proposed model depend on the quality and quantity of the patient observed features. The more features values are observed for a given patient in his EMR and the more diverse patient pool data is, the more accurate readmission prediction would be.

The current model provides a single aggregated latent state. Future work includes expanding the model to create of different latent states, one latent state for each body subsystem. Our final goal is to obtain a global prediction estimate

Discussion

In this paper we have proposed a dynamic model that combines heterogeneous data form the patient Electronic Medical Records to predict the patient readmission to the ICU. The accurate prediction of patient readmission would lead to an good allocation of resources that potentially could reduce cost in the transfer and patient care.

The use of an aggregated patient state which combines current features with previously observed ones allows us to predict the probability of readmission even if not new features are observed. The results of the proposed model depend on the quality and quantity of the patient observed features. The more features values are observed for a given patient in his EMR and the more diverse patient pool data is, the more accurate readmission prediction would be.

The current model provides a single aggregated latent state. Future work includes expanding the model to create of different latent states, one latent state for each body subsystem. Our final goal is to obtain a global prediction estimate
which results from the combination of the subsystems latent states. In the same manner, we plan to expand the model to accommodate different feature weight vectors, one for each disease group. Our hypothesis is that feature values are different across multiple diseases.

Acknowledgments

This work was partially supported by the NIST grant number 60NANB13D136, by NSF/NIST/UMBC grant number SC-0000015277, by CONACYT grant number 207751, by CITRIS SFP 2011-164, and by CITRIS SFP 2015-325. Authors would like to thank Ashit Talukder for his inputs.

References