Predicting Length of Stay for Obstetric Patients via Electronic Medical Records

Cheng Gao\textsuperscript{a}, Abel N. Kho\textsuperscript{b}, Catherine Ivory\textsuperscript{c}, Sarah Osmundson\textsuperscript{d}, Bradley A. Malin\textsuperscript{a-e}, You Chen\textsuperscript{a}

\textsuperscript{a} Dept. of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN, United States
\textsuperscript{b} Institute for Public Health and Medicine, Northwestern University, Chicago, IL, United States
\textsuperscript{c} School of Nursing, Vanderbilt University, Nashville, TN, United States
\textsuperscript{d} Dept. of Obstetrics and Gynecology, School of Medicine, Vanderbilt University, Nashville, TN, United States
\textsuperscript{e} Dept. of Electrical Engineering & Computer Science, School of Engineering, Vanderbilt University, Nashville, TN, United States

Abstract

Obstetric care refers to the care provided to patients during ante-, intra-, and postpartum periods. Predicting length of stay (LOS) for these patients during their hospitalizations can assist healthcare organizations to allocate hospital resources more effectively and efficiently, which, in turn, can improve maternal care quality and reduce patients’ costs. In this paper, we investigate the extent to which LOS is associated with a patient’s medical history. We introduce a machine learning framework to incorporate a patient’s prior conditions (e.g., diagnostic codes) as features in a predictive model for LOS. We evaluate the framework with one-, two-, and three-year historical billing data in electronic medical records for 9,188 obstetric patients in a large academic medical center. The results show that our framework achieve an average accuracy of 49.3%, which is higher than the baseline accuracy 37.7% (that relies on a patient’s age). Furthermore, the most predictive features were found to be statistically significant discriminative features. These included historical billing codes for normal delivery (indicative of shorter stay) and antepartum hypertension (indicative of longer stay).

Keywords:

length of stay (LOS), electronic medical records (EMRs), random forest, obstetrics, prediction

Introduction

Electronic medical record (EMR) systems have been widely adopted in the United States (US) [2-3] and abroad [1,4]. These systems have enabled a substantial amount of patient-specific data to be captured during the routine practice of healthcare organizations (HCOs) [2-5]. This information is quite heterogeneous, ranging from structured diagnoses, medication regimens, laboratory test results and vital signs to un- or semi-structured clinical narratives. The data stored in EMRs is increasingly being recognized for its ability to support numerous activities, such as clinical decision making [5, 19], patient safety improving [20-21] and discovery-driven biomedical research [2-4].

One of the more challenging healthcare management environments in modern time is the safety of maternity. Over the past several decades, the maternal mortality ratio (MMR) has risen dramatically in the US. MMR has doubled from 7.2 deaths of mothers per 100,000 live births in 1987 to 14 in 2015 [6]. At the same time, obstetric care is the most common and costly type of hospital care for all payers in the US [7-9]. Prediction of the length of stay (LOS) for obstetric patients during their hospitalization can help unit managers and administrators to make decisions about hospital resource allocation and configuration. This allows for obstetric care improvement before, during and after childbirth. Better organized care can reduce the morbidity and mortality of women, as well as newborn babies [8,9], and reduce maternity-related costs. Moreover, the families of obstetric patients who are hospitalized frequently inquire about the expected duration of the hospitalization. The incorporation of an accurate estimate of LOS in counseling discussions may mitigate anxieties over the uncertainty of a hospital stay as well as prepare for discharge to home or elsewhere [10].

Previous research has focused on explaining factors that lead to LOS variation in general. LOS has, for instance, been shown to be influenced by numerous factors, including a patient’s demographics (e.g., age), socioeconomic status (e.g., income, education, and occupation), insurance types (e.g., commercial, private and Medicaid and Medicare) and severity of illnesses [11-13]. It has further been shown to be affiliated with HCO-specific factors, such as physicians’ work efficiency [5,13] and the availability of professional language interpretation services [15,16]. However, the complex relationships between these factors further exacerbates the complexity of LOS prediction. Thus, it is challenging to build LOS prediction models that rely solely on expert knowledge and information ascertained at the time of a patient’s admission to a hospital.

Thus, this paper presents a pilot study on the feasibility of a patient’s historical diagnoses, as documented in an EMR for LOS predictive models. This study is predicated on the hypothesis that LOS is related to a patient’s medical history. To investigate this hypothesis, we study three years worth of historic diagnosis codes (prior to their most recent admission) for patients on an obstetric service at Northwestern Memorial Hospital (NMH) in Chicago, Illinois, USA. Specifically, we extracted EMR data in the form of International Classification of Diseases, ninth revision (ICD-9) [17] codes. We designed a machine learning framework to predict LOS. The results indicate that it predicts LOS within 12 hours with over 10% greater accuracy than baseline models that rely solely on the patient’s demographics at time of admission. In addition, we show that certain ICD-9 codes were statistically significant in their predictive capability, which suggests they are ripe further investigation and transition into clinical decision support.

Methods

Figure 1 provides the EMR data and analytics work-flow adopted for this investigation. First, the ICD-9 codes and LOS for patients are extracted from the EMR. These are subsequently applied to train and test a predictive model. Finally, the most discriminant ICD-9 codes are prioritized and assessed for statistical significance.
The dataset was drawn from the Cerner inpatient EMR system in place at NMH from July 2007 to July 2011. It includes the following patient-specific features: 1) demographics (e.g., age), 2) encounter information (e.g., admission and discharge date), 3) diagnosis (e.g., billing codes) assigned to an encounter, and 4) clinical (e.g., obstetrics) service to which the patient was assigned. In total, there were 9188 inpatients in the dataset with 1849 distinct ICD-9 codes. We consider all inpatients on the obstetric service during 2010 and 2011 for prediction and rely on EMR data between 2007 and 2010 as features for our models.

The LOS for an encounter was calculated as the hourly difference between admission and discharge. We use a patient’s age as a baseline prediction for LOS. Table 1 summarizes the average number of ICD-9 codes for the investigated patients in one-, two- and three-year EMRs, the average age of the investigated patients, and the average LOS for these patients on the obstetric service during 2010 and 2011.

![Figure 1](image-url)  
*Figure 1 – The process by which the LOS predictive model is composed and discriminative features are discovered.*

### Dataset

The LOS for an encounter was calculated as the hourly difference between admission and discharge. We use a patient’s age as a baseline prediction for LOS. Table 1 summarizes the average number of ICD-9 codes for the investigated patients in one-, two- and three-year EMRs, the average age of the investigated patients, and the average LOS for these patients on the obstetric service during 2010 and 2011.

<table>
<thead>
<tr>
<th># of ICD-9 codes</th>
<th>Age</th>
<th>LOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 year</td>
<td>2 years</td>
</tr>
<tr>
<td># of ICD-9 codes</td>
<td>4.3</td>
<td>5.5</td>
</tr>
<tr>
<td>min</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>max</td>
<td>60</td>
<td>80</td>
</tr>
</tbody>
</table>

The distribution of inpatients on LOS is shown in Figure 2. The LOS for most (74%) of obstetric patients are ranging from 48 to 96 hours.

![Figure 2](image-url)  
*Figure 2 – Frequency of LOS for inpatients in the study.*

### Predictive Model

We adopted a random forest model to predict a patient’s LOS according to his historic assigned ICD-9 codes. We rely on a random forest because it is a useful ensemble approach for regression and classification. Specifically, the average LOS from all the trees is used for prediction.

We model the data as a matrix, as shown in Equation (1). Let \( n \) be the number of patients and \( m \) be the number of unique ICD-9 codes. In this matrix, each row represents a specific patient and each column is a specific characteristic of the patient. The first column is a patient’s LOS (continuous variable) and the rest of the columns are the ICD-9 codes for each patient. To mitigate the influence of repeat visits for patients, treat each ICD-9 code as a binary variable, such that it is marked as 1 if the patient was assigned this diagnosis at least once and otherwise 0.

\[
\begin{bmatrix}
\text{LOS}_1 & C^1_1 & C^2_1 & \cdots & C^m_1 \\
\text{LOS}_2 & C^1_2 & C^2_2 & \cdots & C^m_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{LOS}_n & C^1_n & C^2_n & \cdots & C^m_n \\
\end{bmatrix}
\]

(1)

To compare performance, we also define a baseline method, which leverages age as a single feature in prediction. Beyond the LOS prediction, the importance ranking of features in the predictive model used will be critical for HCOs to apply and interpret the corresponding results. The random forest regression model enables importance ranking for each feature, which is calculated as extent to which prediction error increases when data for the investigated feature is permuted while all others are held constant [18].

### Performance Evaluation

A random forest regression predicts an LOS as a continuous value. We evaluated the performance of this prediction with respect to the actual LOS as follows. Let us assume \( \text{LOS}_i \) and \( \hat{\text{LOS}}_i \) (\( i = 1, \ldots n \)) are the true and predicted LOS value, respectively. We calculate the difference between \( \hat{\text{LOS}}_i \) and \( \text{LOS}_i \) as \( t_i \). If \( t_i \) is smaller than a predefined tolerance threshold \( \tau \), we claim a correct prediction for the \( i \)-th patient:

\[
b_i = \begin{cases} 
1 & |\hat{\text{LOS}}_i - \text{LOS}_i| \leq \tau \\
0, & \text{otherwise} 
\end{cases} 
\]

(2)

The accuracy (\( \text{Acc} \)) of prediction is thus assessed as

\[
\text{Acc} = \frac{\sum_{i=1}^{n} b_i}{n} 
\]

(3)

### Experiment Design

This section begins with a description of four comparative models in our framework. We then, describe how parameters are selected and compare the models on a range over the parameters. For the parameters selection and models comparisions, we use 5 randomized runs of 3-fold cross-validation. Finally, we describe a statistical test strategy to determine the significant influences of discriminative features on the LOS.

### Comparative Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_0 )</td>
<td>Age</td>
</tr>
<tr>
<td>( M_1 )</td>
<td>One year of ICD-9 codes</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>Two years of ICD-9 codes</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>Three years of ICD-9 codes</td>
</tr>
</tbody>
</table>

Table 2 summaries the four models. The baseline model \( M_0 \) uses age as a lone feature. The other three predictive models (\( M_1 \) to \( M_3 \)) rely on one, two, and three-years worth of historical ICD-9 codes.

### Parameter Selection

There are three parameters in the framework that need to be tuned: 1) number of trees in the random forest, 2) \( \tau \) and 3)
number of ICD-9 codes in the model. Since the models $M_1, M_2$ and $M_3$ are the same in terms of the prediction algorithm, we leverage $M_3$ as a representative model to select an optimal value for each of the parameters.

**Number of Trees**

It has been shown that the performance of random forest regression is relatively insensitive to the number of trees [18]. However, we need to avoid selecting too small of a value (that results in poor accuracy) and too large of a value (that leads to excessive computational load). Thus, we evaluated the number of trees for the random forest. We leverage the distribution of predictive performance to select an optimal number. The model accuracy will grow with the number of trees, up to a certain point, which is where we fix the value.

**LOS Threshold $\tau$**

The threshold, $\tau$ introduced above represents the difference tolerance between predicted LOS and the true LOS and it should be different between HCOs. In this case, the accuracy in LOS prediction is evaluated under a set of thresholds $\{5, 12, 24, 36, 48\}$ satisfy the requirements of disparate HCOs.

**Number of Predictors**

The number of ICD-9 codes (i.e., features) in this study is relatively large (i.e., 1849 in total). As such, we aim to perform dimensionality reduction and derive more manageable model. The specific procedure is as follows. First, we sort all of the predictors on their importance in descending order. Second, we select a subset of the features and predict LOS. We choose the subset with the smallest size but highest predicting accuracy.

**Model Evaluation**

We applied three strategies to evaluate the models. These are summarized in Table 3.

<table>
<thead>
<tr>
<th>Strategy</th>
<th># Trees</th>
<th>$\tau$</th>
<th>% Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Vary</td>
<td>Constant</td>
<td>Constant</td>
</tr>
<tr>
<td>B</td>
<td>Constant</td>
<td>Vary</td>
<td>Constant</td>
</tr>
<tr>
<td>C</td>
<td>Constant</td>
<td>Constant</td>
<td>vary</td>
</tr>
</tbody>
</table>

**Feature Discrimination Analysis**

Two sample t test is used to compare LOS between patients with and without a certain ICD-9 code. We conduct such analysis for the top ten ICD-9 codes with the highest importances derived from the random forest regression models. For each investigated ICD-9 code, we will test the significance of the differences on LOS for patients with and without the code. P value for each significance test will be provided.

**Results**

**Model Paramaterization**

Figure 3 depicts the predictive performance of the models on a varying number of trees. The performance was evaluated while varying $\tau$ over 5, 12, 24, 36 and 48 hours. It was observed that the accuracy did not improve when the number of trees increased beyond 50, which is where we fix the number of trees in the random forest.

The selection of $\tau$ depends on the requirements of an HCOs. It can be seen in Figure 3 that the accuracy grows with $\tau$. If an HCO can accept the predicted LOS between a range of 12 hours shorter and 12 hours longer than the real LOS, then 12 hours could be selected as the value of $\tau$. To this analysis, we set $\tau$ to be 12 hours.

As shown in Figure 4 model accuracy reaches its highest level when the model is based on the top 10% of the features. After this point, accuracy remains relatively constant. As such, our models are based on the top 10% of the features.

**Model Evaluation**

Figures 5-7 depict model performance as a function of the number of trees, LOS threshold and percent of features, respectively. In general, it can be seen that the models that incorporate ICD-9 codes have better performance of LOS than the baseline models. Additionally, models $M_1$, $M_2$ and $M_3$ have almost the same predictive performance. This implies that one-year of historical ICD-9 codes may be adequate for LOS prediction. The accuracy of the four models (where $\tau$ is set to 12 hours and the feature set is fixed to the top 10%) on a varying number of trees (10, 50, 100, 200 and 300) is shown in Figure 5. It can be seen that the accuracy of $M_1$, $M_2$ and $M_3$ is much higher than $M_0$. It can further be seen that the number of trees has minimal influence on the accuracy of all four models. In Figure 6, it can be seen that as the LOS threshold increases, the accuracy of all four models improves. However, the predictive performance does not improve when varying the percent of features (as shown in Figure 7).
outperforms \( M_0 \) (37.7%). Moreover, for patients with LOS \( \geq 96 \) hours, \( M_3 \) outperforms \( M_0 \) as well.

### Table 4 – Model comparison for all patients and patients with LOS \( \geq 96 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>All patients</th>
<th>LOS ( \geq 96 ) hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_0 )</td>
<td>37.7%</td>
<td>0.00%</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>49.3%</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

### Feature Discriminant Analysis

Out of 1849 ICD-9 codes, 9 were selected for further analysis in terms of their clinical implication (as shown in Table 5). Each of the codes in the top 10 ranks exhibited a statistically significant influence on the LOS between LOS for patients with and without such codes. Such evidence may help healthcare organizations (HCOs) to adopt resource allocation strategies that optimize the care management and improve care quality.

As an example, the most discriminant ICD-9 code was 650: “Normal Delivery”. The mean LOS for patients with and without Normal Delivery in their history was 58.1 and 76.5 hours, respectively. The p-value for this difference was less than 0.001. As another example, patients with ICD-9 code of 659.63 “Elderly multigravida with antepartum condition or complication” stayed about 10 more hours than those without that code (81 versus 70.6).

### Discussion

The results show that the hisorical information in EMRs may assist in forecasting obstetric patients’ LOS in a hospital. Specially, our random forest regression model predicted LOS with an accuracy of 49% under an error range of 12 hours, which is 10% more accurate than a baseline mode. The experimental results further demonstrated that a model based on the top 10% of ICD-9 codes can achieve an accuracy as high as those based on all involved ICD-9 codes (over 1800). Additionally, the models based on the most recent year of data logged in the EMRs can achieve the similar performance with those based on two-, and three-year worth of data.

Notably, we also investigated the top 9 ICD-9 codes, which have significant differences in terms of LOS between patients with and without such codes. These results suggest that the HCOs can specialize resource allocation strategies accordingly.

Despite the merits of this investigation, we acknowledge that this is a pilot study and there are several limitations. First, the data was collected within a single institution and may not cover all of a patient’s medical history or be readily applicable to another hospital setting. Second, all of the patients in this study were on an obstetric service, such that the framework may not be directly extended to other types of patients or healthcare services. Third, the prediction of LOS may be considered poor (accuracy of 48% in 12 hours tolerance). Other factors that potentially influences LOS can be incorporated in the model such as certain patient demographics (e.g., race) or physical traits (e.g., height, weight or BMI). Last but not least, if the HCOs can only accept LOS threshold less than 12 hours, the model built in this study will fail.

### Table 5 – Summary for the top 9 most predictive ICD-9 codes

<table>
<thead>
<tr>
<th>Importance Rank</th>
<th>ICD-9 code</th>
<th>Description</th>
<th># of patients (with code verse. without code)</th>
<th>Mean LOS (with code verse. without code)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>650</td>
<td>Normal delivery</td>
<td>2281 vs. 6907</td>
<td>58.1 vs. 76.5</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Conclusions

This paper assessed the feasibility of a machine learning-based framework for predicting LOS for obstetric patients using historical diagnoses. We showed that one year worth of diagnostic history can predict hospitalization LOS with accuracy higher than that of a simple baseline. We plan to extend the framework by including additional types of historical information (e.g., medications) and leveraging the chronological order of such information.

References


Address for correspondence
Cheng Gao: cheng.gao@vanderbilt.edu