# Length of stay prediction for diabetic patients in 130 US hospitals from 1999 - 2008

Yuyu(Ruby) Chen & Siyuan Dong

5/10/2021

## Introduction & Related Work

Diabetes is a serious condition that causes higher than normal blood sugar levels[1]. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths[2]. The inpatient experience and economy burden of diabetes patients has been a raising topic. The largest components of medical expenditures are hospital inpatient care (43% of the total medical cost), prescription medications to treat the complications of diabetes (18%), antidiabetic agents and diabetes supplies (12%), physician office visits (9%), and nursing/residential facility stays (8%). [3] People with diagnosed diabetes, on average, have medical expenditures approximately 2.3 times higher than what expenditures would be in the absence of diabetes. The in-patient cost accounts for a large part of the medical costs for diabetic patients, and it is highly related with length of stay (LOS) in hospital.

Previous articles suggested deep learning techniques such as meta-learning algorithm can be used to have a reasonable estimate on LOS for patients with diabetes, which can help in optimizing the use of hospital resources, reducing healthcare cost, and improving diabetic patient satisfaction. In this study the LOS prediction is explored with different classification techniques using features including age, A1c, glucose level, dose usage, etc.

## Methods

### Data Description

The dataset used for this project is “Diabetes 130-US hospitals for years 1999-2008 Data Set”, which includes the demographic and medical information for 101766 individual diabetic inpatient encounters at 130 US hospitals and integrated delivery networks.

### Statistical Analysis

All analyses were performed using R ver. 4.0.5. For the outcome of interest, length of stay (LOS) was categorized into a dichotomous variable using the average as the cut-off point ($\mu$ = 4.4). 80% of the original data were used as the training set and 20% were used as the testing set with a random split.

Lasso regression and best subset selection were used to select the features for the logistic regression model.

For classification, logistic regression, linear discriminant analysis (LDA), k-nearest neighbors(K-NN) and random forest were used for model fitting and outcome prediction.

For model evaluation, accuracy and AUC were estimated for the 4 models we fit to the data. And the accuracy for the logistic regression model was calculated using a threshold of 0.5.

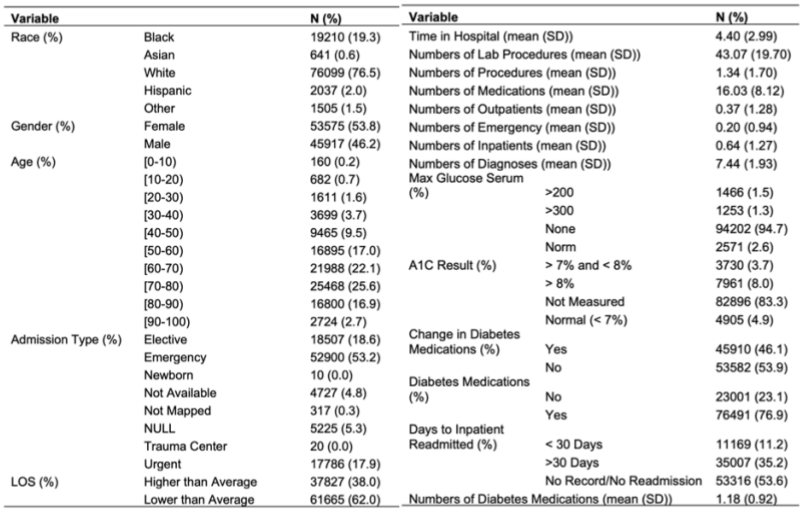
## Results

### Data and Experiment Set Up

The detailed distribution of available variables of this dataset can be found in Table 1. The original dataset contains 24 features for diabetic medications using the generic names, and we summed up the medications prescribed and got the number of diabetic medications administered during each encounter. Then the 24 features were dropped from the dataset. After that, those with unknown gender and race were removed from analysis, which resulted in a final sample size of 99,492.

As seen from Table 1, white (76%), female (54%), and admitted through emergency (53%) account for the majority of the sample. The LOS ranges from 1 to 14 days, with a mean of 4.4 days, and 62% of the patients have a LOS higher than average. The prevalence of lab procedures is low among diabetic inpatients in this sample, where 95% of them did not receive glucose serum test and 83% of them did not receive A1c test.

**Table 1. Descriptive Statistics of Patients' Demographic and Medication (N =99492)**



| Variable | | N (%) | Variable | | N (%) |
|---|---|---|---|---|---|
| Race (%) | Black | 19210 (19.3) | Time in Hospital (mean (SD)) | | 4.40 (2.99) |
| | Asian | 641 (0.6) | Numbers of Lab Procedures (mean (SD)) | | 43.07 (19.70) |
| | White | 76099 (76.5) | Numbers of Procedures (mean (SD)) | | 1.34 (1.70) |
| | Hispanic | 2037 (2.0) | Numbers of Medications (mean (SD)) | | 16.03 (8.12) |
| | Other | 1505 (1.5) | Numbers of Outpatients (mean (SD)) | | 0.37 (1.28) |
| Gender (%) | Female | 53575 (53.8) | Numbers of Emergency (mean (SD)) | | 0.20 (0.94) |
| | Male | 45917 (46.2) | Numbers of Inpatients (mean (SD)) | | 0.64 (1.27) |
| Age (%) | [0-10) | 160 (0.2) | Numbers of Diagnoses (mean (SD)) | | 7.44 (1.93) |
| | [10-20) | 682 (0.7) | | | |
| | [20-30) | 1611 (1.6) | Max Glucose Serum (%) | >200 | 1466 (1.5) |
| | [30-40) | 3699 (3.7) | | >300 | 1253 (1.3) |
| | [40-50) | 9465 (9.5) | | None | 94202 (94.7) |
| | [50-60) | 16895 (17.0) | | Norm | 2571 (2.6) |
| | [60-70) | 21988 (22.1) | A1C Result (%) | > 7% and < 8% | 3730 (3.7) |
| | [70-80) | 25468 (25.6) | | > 8% | 7961 (8.0) |
| | [80-90) | 16800 (16.9) | | Not Measured | 82896 (83.3) |
| | [90-100) | 2724 (2.7) | | Normal (< 7%) | 4905 (4.9) |
| Admission Type (%) | Elective | 18507 (18.6) | | | |
| | Emergency | 52900 (53.2) | Change in Diabetes Medications (%) | Yes | 45910 (46.1) |
| | Newborn | 10 (0.0) | | No | 53582 (53.9) |
| | Not Available | 4727 (4.8) | | | |
| | Not Mapped | 317 (0.3) | Diabetes Medications (%) | No | 23001 (23.1) |
| | NULL | 5225 (5.3) | | Yes | 76491 (76.9) |
| | Trauma Center | 20 (0.0) | Days to Inpatient Readmitted (%) | < 30 Days | 11169 (11.2) |
| | Urgent | 17786 (17.9) | | >30 Days | 35007 (35.2) |
| LOS (%) | Higher than Average | 37827 (38.0) | | No Record/No Readmission | 53316 (53.6) |
| | Lower than Average | 61665 (62.0) | Numbers of Diabetes Medications (mean (SD)) | | 1.18 (0.92) |

## Analysis Results

Lasso regression results suggest that none of the coefficients shrink to 0 (Figure 1). The best subset selection method results display Cp, BIC and adjusted R-squared of the model by the number of variables (Figure 2). All three figures show that when the number of variables is 17, the Cp and BIC are the smallest and the adjusted R-squared is the largest. Both Lasso and best subset selection suggest that none of the variables should be removed from the model.

**Figure 1. The Lasso Coefficient Estimate for LOS (A) and Cross-Validation Binomial Deviance for the Lasso Regression of LOS (B)**

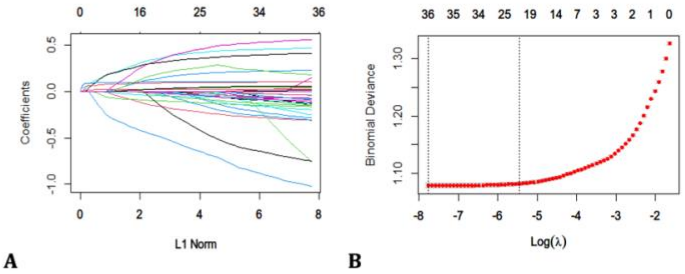
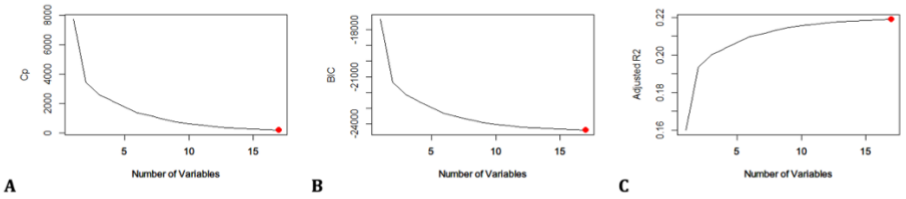


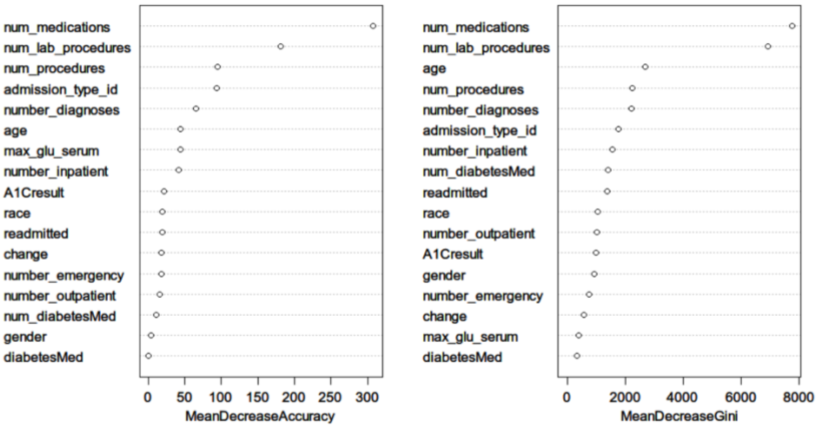
**Figure 2. Cp vs Numbers of Variables (A) BIC vs Numbers of Variables (B) Adjusted R² vs Numbers of Variables (C)**



Among all four models fit the dataset, Logistic regression has the accuracy of 72.67% and AUC of 0.77; LDA has the accuracy of 72.64% and AUC of 0.77; K-NN with k = 29 has the accuracy of 72.22% and AUC of 0.38; and the random forest has the accuracy of 73.36% and AUC of 0.78.

According to the results of model evaluation, the random forest model was the optimal model for prediction, which has the largest accuracy and AUC. This means that the random forest model could classify 73.36% of the truth, and the probability that a randomly selected patient with LOS greater than average will rank higher than a randomly selected patient with LOS less than average for the random forest model is 0.78, which suggests a good performance of this model. As seen from Figure 3, the number of medications taken and the number of lab procedures performed for diabetic patients have higher importance in the random forest model, where removing these two variables out will lead to a much greater decrease in mean accuracy and Gini index compared to removing other variables out.

**Figure 3. Random Forest Variable Importance plot of LOS**



## Discussion

This study predicts whether the LOS for diabetic inpatients is above or below average using the demographic and medical information. The number of medications administered and the number of lab procedures performed during the encounter are two most important features for predicting the LOS. The number of medications reflects the severity and complexity of the diseases for that patient, where those who take more medication could have worse health conditions. Also, medications administered could include those other than diabetic medications. Higher number of medications administered might indicate the patient has a higher chance of having diabetes-related complications, such as nonfatal stroke and nonfatal ischemic heart disease, which are the most common seen complications for hospitalizations.[4] The number of lab procedures also reveals the health status for the patients, where more lab procedures performed during the encounter indicates more severe and complex illness. The patients with more serious conditions are more likely to stay longer in hospital. In addition, more lab procedures performed will take more time, which could also result in a longer LOS.

The limitation of this study is that there might be some loss of continuous variation by categorizing LOS based on the average. Although the average LOS is a well-established measure for determining the severity of inpatients, the actual number of days stayed might contain more information about the health condition. Future work may involve treating LOS as a continuous outcome using machine learning algorithms.

**Reference:**

1. Diabetes - Symptoms and causes - Mayo Clinic. Accessed May 10, 2021. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

2. Diabetes. Accessed May 10, 2021. https://www.who.int/news-room/fact-sheets/detail/diabetes

3. Economic Costs of Diabetes in the U.S. in 2012 | Diabetes Care. Accessed May 10, 2021. https://care.diabetesjournals.org/content/36/4/1033.short

4. Cheng S-W, Wang C-Y, Ko Y. Costs and Length of Stay of Hospitalizations due to Diabetes-Related Complications. *Journal of Diabetes Research*. 2019;2019:e2363292. doi:10.1155/2019/2363292

**Contributions:**

Yuyu(Ruby) Chen: Model fitting and the methods and results accordingly, Introduction and Related work Section, table for the data and experiment setup

Siyuan Dong: Model evaluation and the methods and results accordingly, Discussion Section, write-up for the data and experiment setup