A Generalized Linear Mixed Model for Readmission Prediction among Patients with Diabetes

Wang Zhenyu

August 2020

1 Introduction

Along with the wide spread of the COVID-19 epidemic, the healthcare system overloads and hovers on the brink of collapse. As the most common disease in the world, diabetes, also contribute to this situation. Patients with diabetes perform bad outcomes even after medical treatment, so some of these patients readmit multiple times and are more costly to the healthcare system. In order to reduce the unnecessary burden on the system, it is important to classify the patients who tend to have bad result and provide them targeted medical treatment to improve their end results and lower the overall expenditure.

Readmission (to hospitals) may due to inadequate care of the hospital or weak health condition of the patient. Hence it is a suitable measurement for the cost of healthcare. Therefore, the goal of this project is to develop a model to predict the readmission of with diabetes patients based on **diabetes** dataset, describe associations between readmissions and other medical characteristics, and give some suggestions on healthcare system.

2 Methods

2.1 Data Preparation

The raw **diabetes** dataset contains 101766 observations from 71518 patient. Each observation is an encounter of a patient. There are 48 potential covariates can be used to predict *Readmitted* variables. These covariates are related to encounter's treatment record, patient's personal information, characteristics, medication features. More exploratory data analysis are described in result section.

The purpose of the study is to predict readmission of an encounter and do not discuss about the development of disease of a patient. Due to complexity of the data and the purpose of this study, the Generalized Linear Model (hereinafter called the 'GLM') is applied to this analysis (Model selection will be discussed in the 4.2 part). The data was collected through longitudinal study, and it leads to within-subject correlation. That is, encounters within each patient are dependent. This dependence violates one assumption of GLM, that is, every observation has to be independent. To fix this, only the first encounter for every patient is selected for further analysis, and there are 71518 observations left. Some categorical covariates should be recategorized to reduce complexity of the model. For example, some medication features variables have four level to describe ('Steady', 'Up', 'Down', 'No'), and more than 95% observations have 'No' value, combining rest three variables as 'Yes' variable not only satisfies its contextual meaning, but also makes model easier to construct and interpret. Log transformation may apply to some numeric variables in order to scale down the data.

For the sake of model validation, 20000 observations (encounter) are randomly chosen to be testing data set, the rest are used to train the model.

2.2 Data Analysis

Since the study only predict whether or not readmission happens, the response variable *readmitted* is recategorized into a binary variable with Yes (previous $>$ 30 and $<$ 30) and No values. Thus the response variable is binary, that is, an independent Bernoulli trail, a logit link function is applied to GLM.

After construction of the full model, model diagnostics are performed to verify logistic model assumptions: the Debeta Plot for influential observations removal, the Deviance Residual Plot for independence assumption. If there are existence of bad influential points, we may remove them if there is a good contextual justification or keep it as a potential limitation. If independence assumption is violated after 2.1, then review on data collection and design of experiment is needed.

After model assumptions are satisfied, step methods of AIC-criteria, BIC-criteria and elastic-net method are applied for variable selection. Using three methods, we obtain three candidate model, and they will be compared through AIC, BIC and other contexts. The model with lower AIC, BIC value is preferred, other than this, the model with less covariates and easier to interpret is preferred.

Then, add interaction terms into the model, interaction terms are always between patient's person characteristics (e.g. age, gender...) and medical characteristics (e.g. insulin, Number of Medication...) in this context. Likelihood ratio test will be applied to compare nested models and justify variable removal because of contextual reasons or interaction terms. Variable selection may go back and forth till a good model is obtained with appropriate number of predictors (preferably below 10), and good performance based on AIC, BIC values.

Once a final model is obtained, the cross validation method on the whole dataset is applied to validate model, and calibration plot can be used to access accuracy of prediction. Also, run the model on testing dataset, see if significance and coefficents of variables are similar for model on testing and training sets. Lastly, use ROC curve to determine the performance of the model on the testing data set. Higher Area Under Curve (hereinafter called the 'AUC') indicates better discriminability. Also, compare the AUC between model prediction on training data set and testing data set. If the results are similar, indicates the model has a good fit and is validated.

The regression analysis is conducted in R (Version 4.0.0).

3 Results

3.1 Description of Data

This describes data after preparation in part 2.1. Characteristics of numeric variables are shown in Figure 1 through histograms. Most encounters more than 4 number of diagnoses, while many of them do not have emergency or outpatient procedures (89.9%, 83.7%). More encounters have *lengh of stay* within 5 days (Mean = 4.395987, Median = 4).

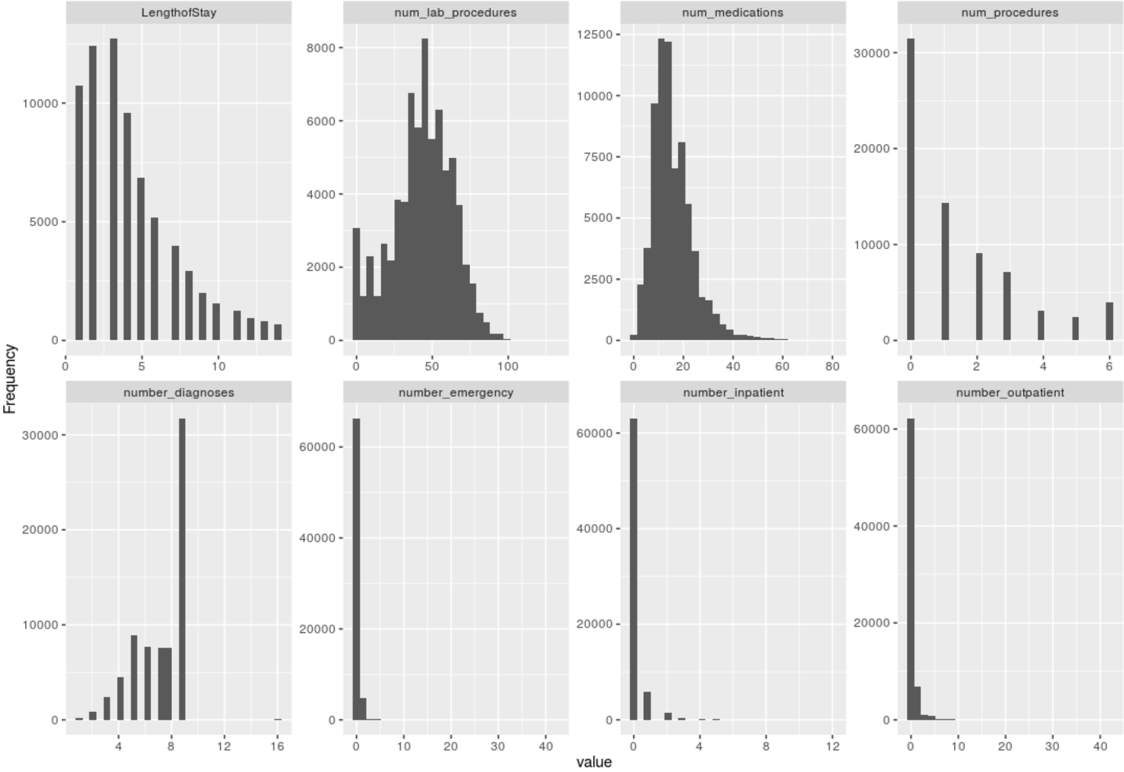


Figure 1: Numeric Variable Summary

Characteristics of numeric variables are shown in Figure 2 through bar plots. The response variable *readmitted* has similar proportions of YES (49.9%) and NO value (60.1%). As for *race*, most of them in this sample are Caucasian. The majority of encounters have No values in all medical feature variables. *diabetesMed*, *examide* and *citoglopton* even do not have Yes values among all patients.

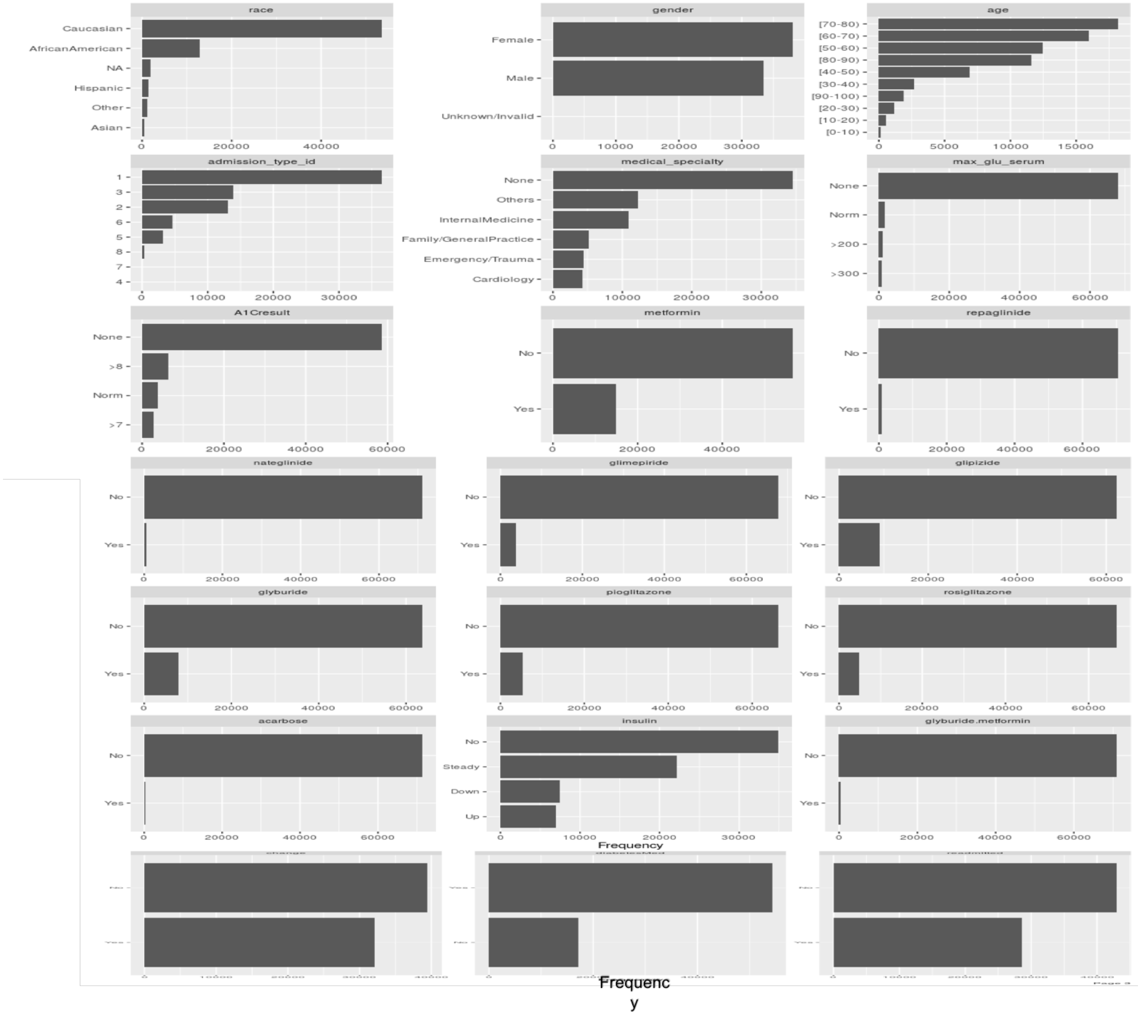


Figure 2: Categorical Variable Summary

Transformation or recategorization of variables are recorded Table 1. Dropped variables and reasons (large missing value, sparse variables, etc.) for dropping are recorded in Table 1, and will be discussed more in 3.3. At this stage, 31 potential predictors left to build the full model, 10 of them are numeric, 21 are categorical.

3.2 Model Diagnostic

There are several bad influential observation, but no reason to drop them. And there is no very obvious evidence of dependence between responses.

3.3 Variable Selection

Before automatic variable selection, some variables have more than 90% missing values. They will be dropped to keep most observations.

Some categorical variables are extreme sparse even after combining variables and recategorization. If a variable has a value in more than 99% observations, they will be dropped, because they affect the construction of the model.

The automatic variable selection using AIC, BIC criteria and Elastic-Net method are recorded in Appendix Table 3. Elastic-Net model is worse because of higher AIC, BIC values. Both two step method selection have a higher value and lower value in AIC and BIC. However, AIC-criteria selected model has 17 covariates (higher than 12 of BIC-criteria), which makes the model hard for data collection and interpretation. Therefore, step method BIC-criteria selected variable is chosen as a base model. Table 1 records every variable's selection procedure.

Since 12 is still considered to be a high number of predictors. Likelihood ratio test is performed to delete specific variables. Admission_type_id variable is dropped use likelihood ratio test ($teststatistic = 9.80325, p - value = 0.1938, df = 7$). Race variable is dropped since it contains 2% missing values, though this proportion is not huge, but this lower total number of covariates to 10.

Add more interaction terms to the model and use likelihood ratio test to check significance of newly added terms. The result of likelihood ratio test for nested model comparison are generally insignificant, there is no evidence to conclude difference between a model contains interaction term and a model not contains. Therefore, no interaction terms are necessary to added in the final model.

Variable	Selection	Transformation	Drop Reason
encounter_id	N	N	ID
patient_nbr	N	N	ID
race	A	N	Context
gender	AE	N	Auto
age	ABE	N	Final
weight	N	N	Missing
admission_type_id	BE	N	Context
discharge_disposition_id	N	N	Sparse
admission_source_id	N	N	ID
Length of Stay	ABE	N	Final
payer_code	N	N	ID
medical_specialty	AE	Recategorization	Auto
num_medications	N	N	Auto
number_outpatient	ABE	N	Final

number_emergency	ABE	N	Final
number_inpatient	ABE	N	Final
number_diagnoses	ABE	N	Final
max_glu_serum	N	N	Auto
A1Cresult	AE	N	Auto
metformin	B	N	Final
repaglinide	N	Recategorization	Auto
nateglinide	N	Recategorization	Auto
chlorpropamide	N	N	Sparse
glimepiride	N	Recategorization	Auto
acetohexamide	N	N	Sparse
glipizide	E	Recategorization	Auto
glyburide	N	Recategorization	Auto
tolbutamide	N	N	Sparse
pioglitazone	E	Recategorization	Auto
rosiglitazone	E	Recategorization	Auto
acarbose	N	Recategorization	Auto
miglitol	N	N	Sparse
troglitazone	N	Recategorization	Sparse
tolazamide	N	N	Sparse
examide	N	N	Sparse
citoglipton	N	N	Sparse
insulin	ABE	N	Final
glyburide.metformin	N	Recategorization	Auto
glipizide.metformin	N	N	Sparse
glimepiride.pioglitazone	N	N	Sparse
metformin.rosiglitazone	N	N	Sparse
metformin.pioglitazone	N	N	Sparse
change	AE	N	Auto
diabetesMed	ABE	N	Final
readmitted	N	Recategorization	Response
encounter_num	N	N	ID

Note: In selection column, letter “A” indicates this variable is selected by AIC-criteria step method.
Letter “B” indicates this variable is selected by BIC-criteria step method.
Letter “E” indicates elastic-net method selected variable.
Letter “N” indicates no method selects this variable.
In transformation, Letter “N” indicates no transformation is applied.
In Drop Reason Column, “ID” indicates this variable is dropped due to it is an ID variable.
“Missing” indicates this variable is dropped due to large proportion of missing value.

"Auto" indicates this variable is dropped in the automatic variable selection procedure.

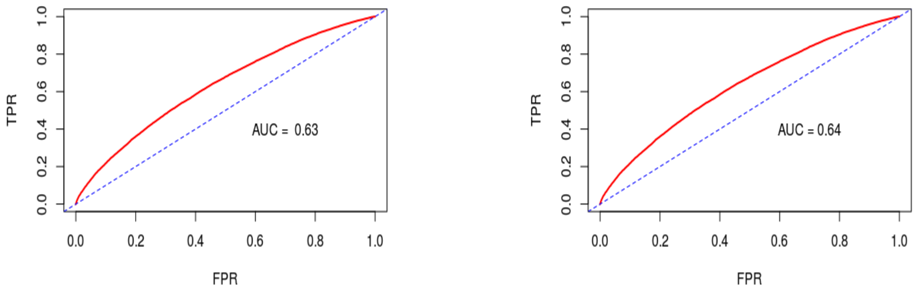
"Sparse" means this variable may be too concentrated for one value, and extreme sparse for other values.

Table 1: Variable Selection Result

3.4 Goodness of Model

Use the model selected in 3.3 (result in Table 2), use cross-validation method on the whole dataset. In the calibration plot(Appendix, figure 4), bias-corrected line fits ideal line without big differ and prediction accuracy is acceptable.

Run the model on testing dataset. The ROC curves of this model on training and testing datasets is ploted in figure 3. On the testing dataset, the AUC is 0.63, which indicates that this model has a reasonable discrimination ability (It can discriminate he model can correctly discriminate between the readmission or not 63% of the times.) Also, comparing AUC of model on two datasets, AUC for training dataset is 0.64, 0.01 is trivial difference, thus, we may conclude this model does not overfit the training data.

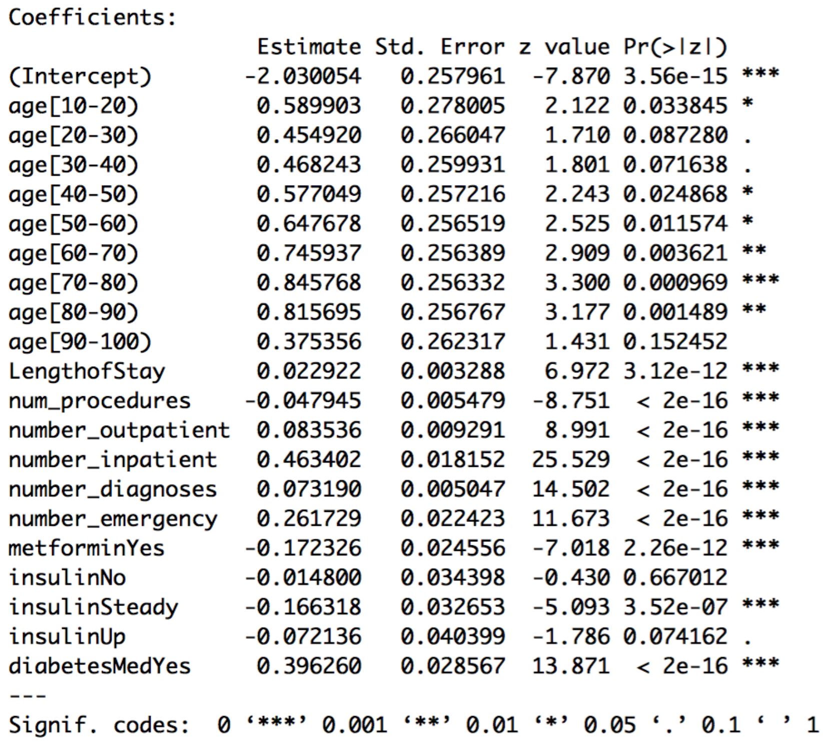


(a) Testing Dataset

(b) Training Dataset

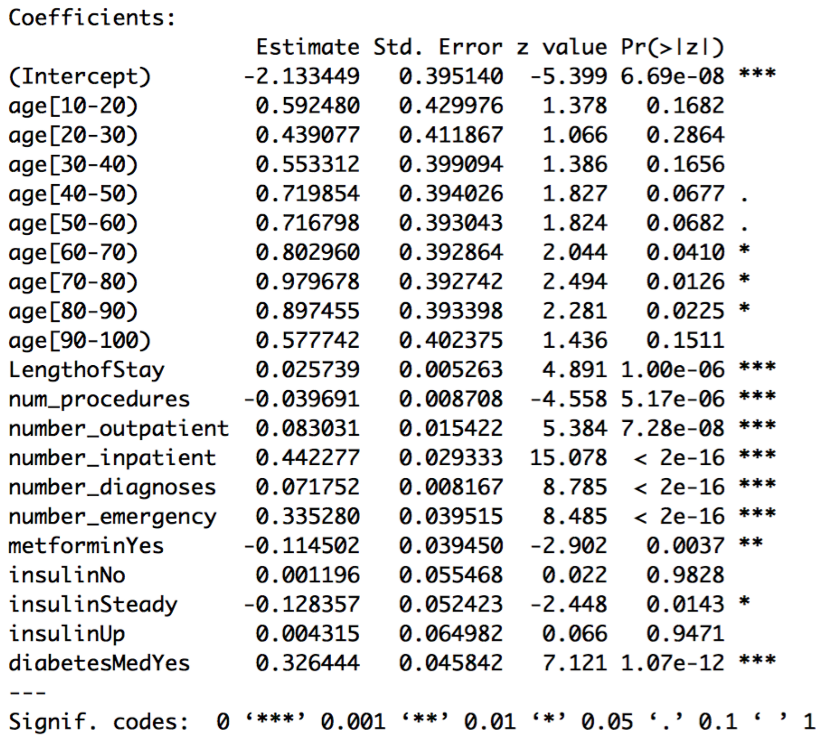
Figure 3: The ROC Curve for the final model

The coefficients and significance of variables of this model are recorded in the appendix. We see that despite *age* variable, two datasets give similar results in general. This confirms this model does not overfit, and the performance of this model is validated.



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -2.030054   0.257961  -7.870 3.56e-15 ***
age[10-20)         0.589903   0.278005   2.122 0.033845 *  
age[20-30)         0.454920   0.266047   1.710 0.087280 .  
age[30-40)         0.468243   0.259931   1.801 0.071638 .  
age[40-50)         0.577049   0.257216   2.243 0.024868 *  
age[50-60)         0.647678   0.256519   2.525 0.011574 *  
age[60-70)         0.745937   0.256389   2.909 0.003621 ** 
age[70-80)         0.845768   0.256332   3.300 0.000969 ***
age[80-90)         0.815695   0.256767   3.177 0.001489 ** 
age[90-100)        0.375356   0.262317   1.431 0.152452    
LengthofStay       0.022922   0.003288   6.972 3.12e-12 ***
num_procedures    -0.047945   0.005479  -8.751  < 2e-16 ***
number_outpatient  0.083536   0.009291   8.991  < 2e-16 ***
number_inpatient   0.463402   0.018152  25.529  < 2e-16 ***
number_diagnoses   0.073190   0.005047  14.502  < 2e-16 ***
number_emergency   0.261729   0.022423  11.673  < 2e-16 ***
metforminYes      -0.172326   0.024556  -7.018 2.26e-12 ***
insulinNo         -0.014800   0.034398  -0.430 0.667012    
insulinSteady     -0.166318   0.032653  -5.093 3.52e-07 ***
insulinUp         -0.072136   0.040399  -1.786 0.074162 .  
diabetesMedYes     0.396260   0.028567  13.871  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Result on Testing Dataset



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -2.133449   0.395140  -5.399 6.69e-08 ***
age[10-20)         0.592480   0.429976   1.378   0.1682    
age[20-30)         0.439077   0.411867   1.066   0.2864    
age[30-40)         0.553312   0.399094   1.386   0.1656    
age[40-50)         0.719854   0.394026   1.827   0.0677 .  
age[50-60)         0.716798   0.393043   1.824   0.0682 .  
age[60-70)         0.802960   0.392864   2.044   0.0410 *  
age[70-80)         0.979678   0.392742   2.494   0.0126 *  
age[80-90)         0.897455   0.393398   2.281   0.0225 *  
age[90-100)        0.577742   0.402375   1.436   0.1511    
LengthofStay       0.025739   0.005263   4.891 1.00e-06 ***
num_procedures    -0.039691   0.008708  -4.558 5.17e-06 ***
number_outpatient  0.083031   0.015422   5.384 7.28e-08 ***
number_inpatient   0.442277   0.029333  15.078  < 2e-16 ***
number_diagnoses   0.071752   0.008167   8.785  < 2e-16 ***
number_emergency   0.335280   0.039515   8.485  < 2e-16 ***
metforminYes      -0.114502   0.039450  -2.902   0.0037 ** 
insulinNo          0.001196   0.055468   0.022   0.9828    
insulinSteady     -0.128357   0.052423  -2.448   0.0143 *  
insulinUp          0.004315   0.064982   0.066   0.9471    
diabetesMedYes     0.326444   0.045842   7.121 1.07e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Result on Training Dataset

Table 2: Final Model Output

4 Discussion

4.1 Interpretation

From the model result on the training dataset, at significance level of 0.05, the associations between the readmission and age, length of stay, etc. are significant. Use the estimated coefficients, the associations can be quantified in terms of the odds ratio. For instance, for categorical variable *metformin*, when other variables keep fixed, the odds ratio between a patient take metformin and not take metformin is approximately 0.84 (exp(-0.17)). For numeric variable, *num_inpatient*, when other covariates keep constant, one unit increase in number of inpatient for an encounter will lead to approximately 0.58 (exp(-0.46) = 1.58) increase in odds value.

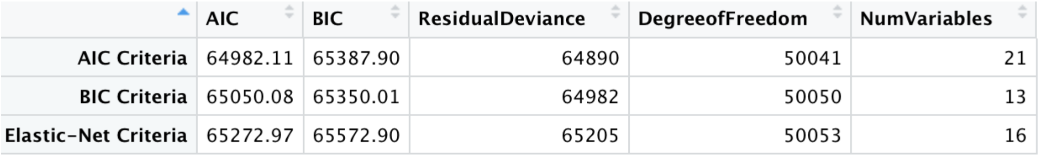
In addition to the numeric results itself, it is easy to see number different types of procedures contribute to the probability of readmission of patients with diabetes, the hospitals are recommended to keep track of this information to put more recourse on those patients with more emergency experience. Also, *insulin* and *metformin* are two medicine covariates, their performance and medication habits affect the readmission probability and furthermore, the overall cost of the health system.

4.2 Limitation

Since this is a longitudinal study, Generalized Linear Mixed Model is more suitable to overcome dependence. However, due to the large scale of the data set and the complexity of the model itself, running GLMM with **Diabetes** on RStudio Cloud is extreme time consuming and may not converge. Therefore, the trade-off in this study is to waste 30248 observations in the original dataset to apply simpler model with a more strict independence assumption. This loss of information leads to a model with less accurate prediction. A machine with better performance can improve the running time of model construction in the future.

This study is strongly based on medical background (for some important variables). Some variables or interaction terms may be very important because of contextual reasons, but are automatically removed. This may not influence the performance of this model, but may reduce interpretability. Therefore, researchers with more medical background should review this study.

Appendix



	AIC	BIC	ResidualDeviance	DegreeofFreedom	NumVariables
AIC Criteria	64982.11	65387.90	64890	50041	21
BIC Criteria	65050.08	65350.01	64982	50050	13
Elastic-Net Criteria	65272.97	65572.90	65205	50053	16

Table 3: Automatic Selection Model Comparison

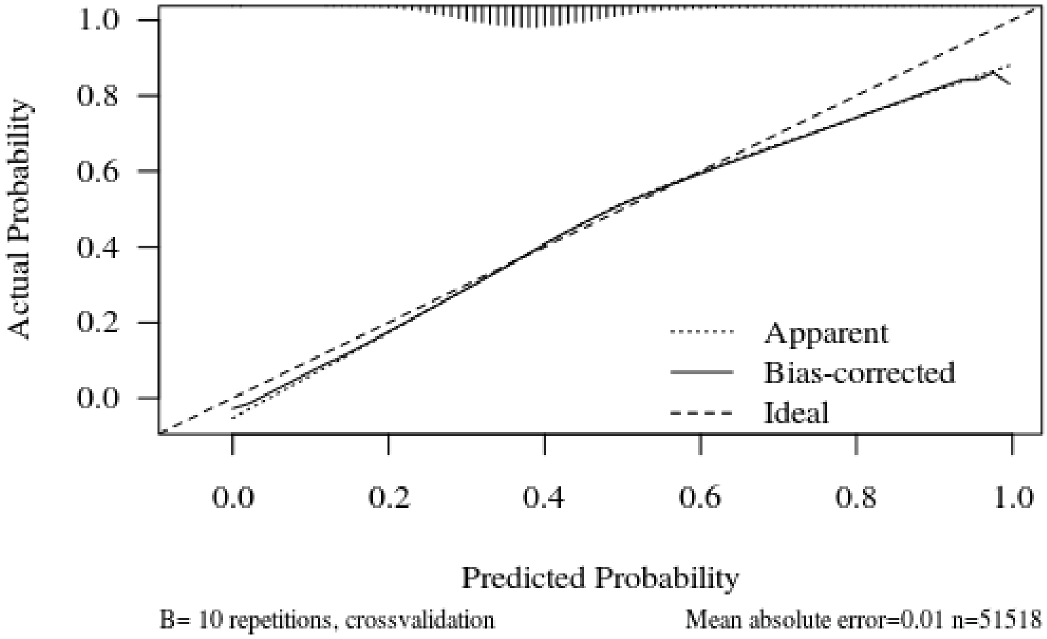


Figure 4: Calibration Plot for Cross Validation