Predicting Readmission in Diabetic Patients

Lisa Chae Young Oh
August 30, 2020

1 Introduction

Diabetes is one of the top chronic diseases the world, with a high prevalence rate and a common cause of death globally. It holds burdens to the people who are affected with this lifelong disease and to the healthcare system supporting them. Health facility patients who suffer from diabetes tend to have worse medical outcomes than non-diabetic counterparts. These patients can be costly to the healthcare system as further treatments or future readmissions are likely. Yet this does not mean all diabetic patients incur higher costs. Therefore, the goal is to identify the characteristics of patients who are more likely to have worse outcomes so early intervention can be implemented to reduce the burden.

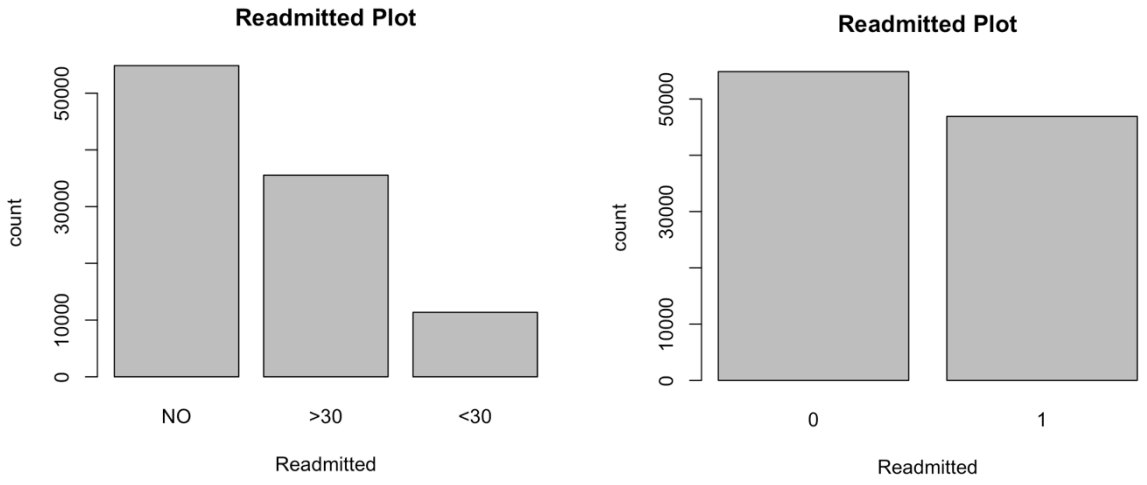
One measure of healthcare burden is hospital readmission. A patient can be readmitted to a hospital within a certain period of time for one of many reasons, including inadequate care on the initial stay and general poor health of the patient. This results in high costs to the healthcare system. This report presents an approach to predicting readmission in diabetic patients based on a range of patient and hospital outcomes.

2 Methods

2.1 Choice of Methods

In the original dataset, the response variable *readmitted* had three categories: no readmission, a readmission in less than 30 days, and readmission in more than 30 days. In this approach, the latter two categories were treated as the same. That is, the response variable was dichotomized to be in one of two categories: no readmission or readmission. As Figure 1 shows, after the dichotomization, the number of patients in the two categories are similar but there still remains more patients who were not readmitted than readmitted. Because the response is a binary categorical variable, logistic regression was chosen to model the problem. Logistic regression also has good interpretability, so explaining model results will be straightforward.

Figure 1: Summary of readmission before vs. after dichotomization



2.2 Variable Selection

Three different models were created from an initial model (consisting of all covariates) by stepwise variable selection by AIC, stepwise variable selection by BIC, and model shrinkage by LASSO. The stepwise methods were applicable since the number of observations was much larger than the number of predictors. The LASSO method performs both regularization and variable selection. With LASSO, there is possibility of overpenalizing the coefficients of the regression variables since the number of predictors is bounded by the sample size, but this is not a concern since the dataset is much larger than the number of predictors. The variables with non-zero coefficients resulting from the LASSO method were kept as part of the model.

2.3 Model Violations/Diagnostics

The logistic regression model assumes independence. The dataset violates this assumption because if a patient has multiple hospital admissions, all of these encounters were recorded. The initial approach was to take the most recent encounter for all patients, however this could lead to loss of valuable information in predicting readmission and assumes that the outcomes of the latest encounter predicts readmission. Looking at the distribution of number of encounters for each unique patient (Figure A.1 in Appendix), it turns out that approximately 76.5% of patients only had one encounter. Since this represents the majority of the dataset, it is assumed that the observations are independent and no changes were made.

Model diagnostics were performed by the Receiver Operating Characteristic curve (ROC curve) and binned residuals. Cross-validation was used for model validation.

3 Results

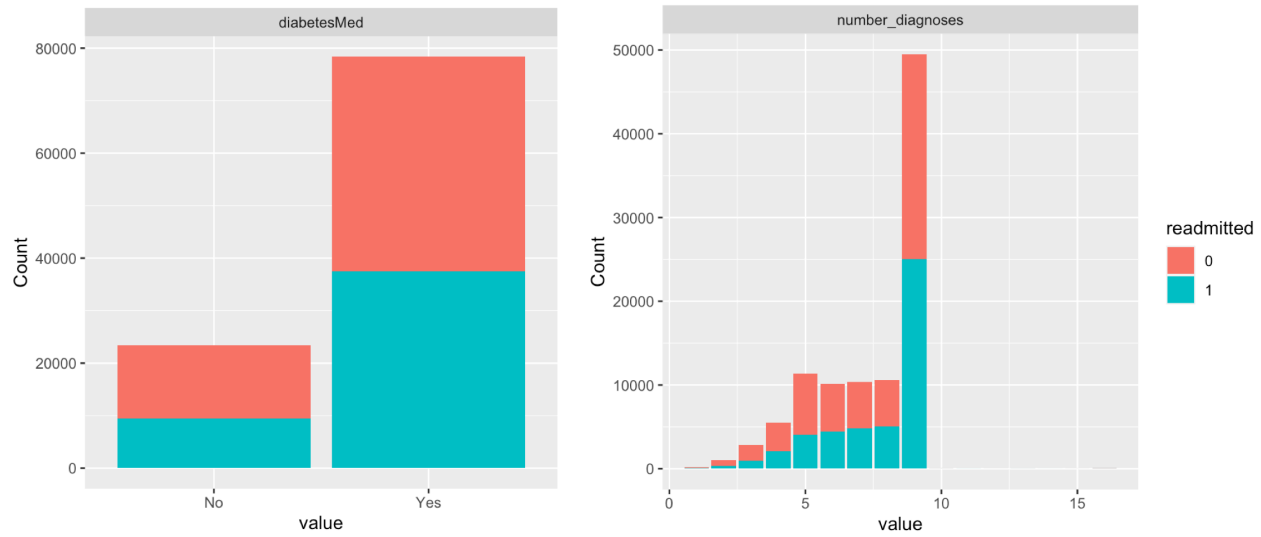
3.1 Description of Data/EDA

Prior to fitting the model, the dataset was explored and cleaned. First, the number of missing values was counted for each variable. If the variable was a potential predictor (meaning not a response variable or identification variable) and had notably many missing values, the variable was removed from the dataset because it would not be a reliable predictor. This resulted in the removal of variables *weight* and *medical_specialty*. Unnecessary identification variables were removed as well.

While examining the distribution of all variables, it was seen that a high proportion of certain diabetes medication variables were categorized as “No” (the medication was not administered), as shown in Appendix figure A.2. Therefore, medication variables were removed if over 95% of encounters had value “No” since they would not be important predictors.

Some interesting variables were *diabetesMed* (whether diabetes medication was prescribed) and *number_diagnoses* (number of diagnoses in current encounter). Figure 2 shows that regardless of whether or not the encounter was a readmission, a significant proportion of encounters had new prescriptions of diabetes medication and a large number of diagnoses.

Figure 2: Summaries of dataset features



Furthermore, the distributions showed that some categories of variables have an extremely low number of observations. Therefore covariates *admission_type_id, discharge_disposition_id, number_emergency, number_inpatient, number_outpatient* were reorganized to have fewer categories.

3.2 Process of Obtaining Final Model

The dataset was split into train and test sets with observations from 51,518 and 20,000 patients respectively. The initial model was fit with all covariates included and the training set as the data. The next step was variable selection since there were many predictors in the initial model.

The models presented by stepwise selection by AIC, stepwise selection by BIC, and LASSO had 19, 13, and 14 predictors respectively. They were compared on their goodness of fit. Some results of these diagnostics are below.

Figure 3: Model diagnostics

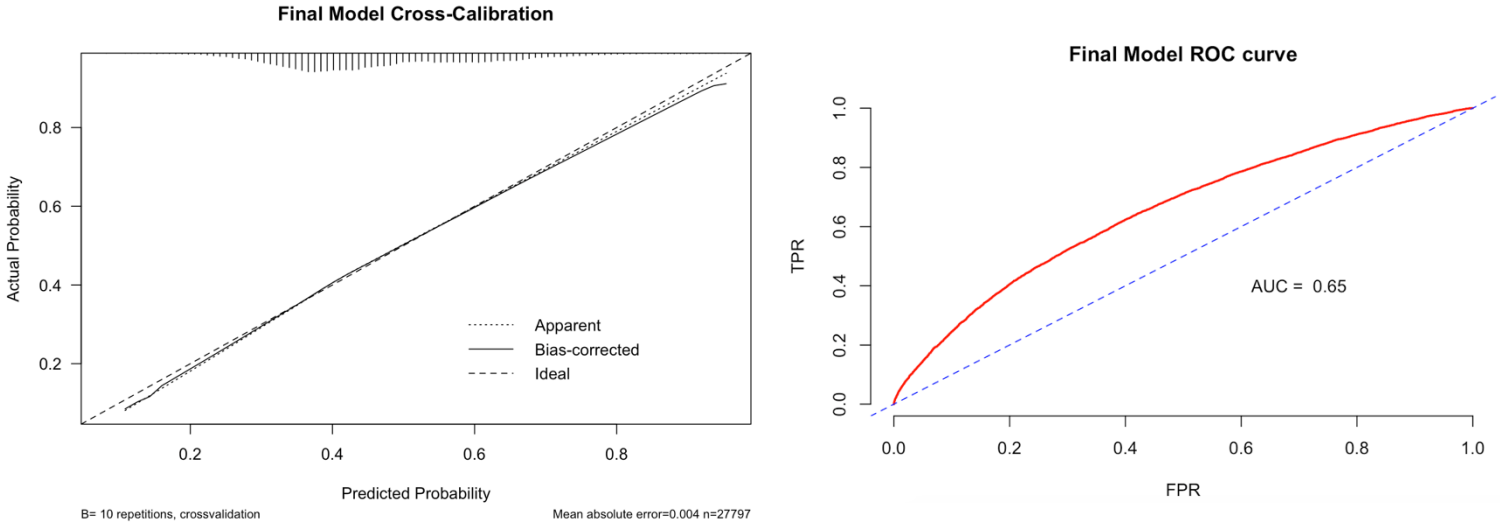
Variable Selection Method	AUC of ROC curve
Stepwise (AIC)	0.6621
Stepwise (BIC)	0.6606
LASSO	0.6609

The models all have nearly identical AUC values. Furthermore, plots of binned residuals against linear predictor and fitted values were very similar for the three models. They were scattered yet depending on the size of bins, a positive relationship could be observed. Cross-calibration plots were also very similar for the three models, exhibiting good validation results. Therefore, the model chosen by the stepwise BIC method was chosen as the final model as it had the least number of predictors.

3.3 Goodness of Final Model

Diagnostic results of the final model's predictions on the test set are presented in Figure 4. The cross-calibration plot for model validation demonstrates that the Bias-corrected line deviates from the Ideal line particularly at the tails. Thus the model performs well in predicting responses from the test dataset. However, the ROC curve close to the 45° line, giving an AUC value of 0.65. This model has a moderate discrimination ability between readmission and non-readmission; perhaps it is randomly deciding readmission values.

Figure 4: Goodness of final model



4 Discussion

4.1 Final Model Interpretation and Importance

It turns out that the variables of interest presented in section 3.1 were included in the final model. Below is a summary of their outputs.

Figure 5: Selection of variables from the final model

Variable	Coefficient	p-value	95% CI
number_diagnoses	0.068740	< 2e-16	(0.0599, 0.0776)
diabetesMedYes	0.326921	< 2e-16	(0.2787, 0.3752)

These two variables are an example of a continuous variable and a categorical variable respectively. They are significant because they have a small p-value and their confidence intervals do not contain zero. *number_diagnoses* represents the number of diagnoses during the encounter. The interpretation is as such: for each additional diagnosis, the odds of readmission increase by a multiplicative factor of exp0.068740)=1.07. Moreover, *diabetesMedYes* is a category for the *diabetesMed* variable representing whether diabetes medication was prescribed in the encounter. The reference category is *diabetesMedNo*. The interpretation is as such: the

odds of readmission is expected to increase by a multiplicative factor of exp(0.326921)=1.39 for encounters with diabetes medication prescriptions as compared to encounters without such prescriptions.

Overall, encounters with higher number of diagnoses present more burden on the healthcare system, and encounters with diabetes medication prescriptions present more burden on the healthcare system than encounters without diabetes medication prescriptions.

4.2 Limitations of Analysis

For variable selection, the stepwise method is problematic when covariates are highly correlated and coefficients may be biased. Hence the resulting set of covariates may not be the best predictors of readmissions. As well, we made the assumption that independence is met even though all encounters of a patient were recorded. Thus a generalized linear mixed model is suggested as an alternative to include patients as a random effect.

This was presented as an approach to modelling the problem. However from lack of medical background and other unrecorded yet potentially important factors from encounters and individual health states, it is not advised to use this model for healthcare recommendations.

5 Reference

National Diabetes Statistics Report, 2017 Estimates of Diabetes and Its Burden in the United States Background. (2020). https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

WHO. (2018, May 24). *The top 10 causes of death*. Who.Int; World Health Organization: WHO. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

6 Appendix

Figure A.1: Distribution of maximum number of encounters for each unique patient

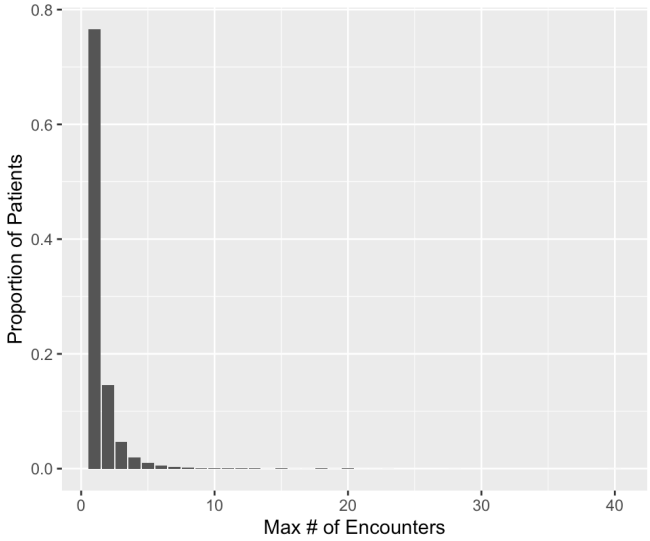


Figure A.2: Excerpt from diabetes medication summary

Diabetes Medication	Category	Proportion of Total
citoglipton	No	100%
examide	No	100%
acetohexamide	No	100%
glimepiride.pioglitazone	No	100%
metformin.pioglitazone	No	100%
metformin.rosiglitazone	No	100%
troglitazone	No	100%
glipizide.metformin	No	100%
tolbutamide	No	100%
glipizide.metformin	No	100%
…	…	…