A Population-based study of Hospital Readmission Risk for American patients

Author: Yuanzhe Yang

2020/12/20

Abstract

Patients who are encountered to hospital within 30 days would be identified as a readmission cases. These unplanned reemissions not only suggest healthy risks caused by insufficient medication and treatment, but also increase the government financial costs. For providing profound insights of readmission risk, we focused on a dataset containing national clinical information data collected from over 130 US hospitals, and investigated several determinants that might be associated with the hospital readmission, such as diabetes, patients' medical history and status, by generalized linear model (GLM) and generalized linear mixed model (GLMM). Both the numerical and graphical evidence implicated that diabetic patient has a greater risk of readmission than others. We also found that the classification by GLMM was more accurate than GLM,

Keywords: Readmission risk, Diabetic patients, GLM, GLMM, ANOVA, ROC curves

Introduction

People's daily lifestyles have been enormously improved since the early 21st century, due to the rapid development of global economy and scientific techniques. However, people are suffering the curse of advanced technology at the same time: improperly managed diet, insufficient physical excursive and other unhealthy behaviors bring people in a risk of common disease: diabetes. Although this chronic disease does not present apparent symptoms, it has no specific medical treatment for instant cure at present and diabetic patients tend to be more sensitive to be involved in other dangerous diseases, for instance, heart disease or pneumonia, due to their vulnerable immune system. In the United States, diabetic patients averagely cost the Medicare extra 250 million dollars spend for their readmitted retreatment since 2011 [1], which definitely increased a financial burden on federal government. On the other hand, the rate of readmitted patients is commonly utilized as a criterion of evaluating the performance of a specific hospital: the poor diagnosis or invalid treatment could be distinguished by the high readmission rate of patients [2]. Consequently, investigate the relationship between hospital readmission and diabetes, and evaluate the influence of potential factors related with the risk of readmission could provide some profound insight for reducing the federal cost while also providing sufficient support to patients with more reasonable arrangement of medical resource.

Generally, several methods and programs had been utilized for controlling repeated hospitalization in recent years. Unplanned readmission risk could be decreased by more tightened discharge rules as proposed by Goudjerkan et al. [3]. The U.S. Centers for Medicare and Medicaid Services (CMS) has conducted an initiative program called Hospital Readmission Reeducation Program (HRRP) since 2012, sufficiently decreased the rate of repeated encounters [4].

In this report, we initiated with our study with data cleaning and preliminary analysis on the dataset from the UCI Machine Learning Repository. The response variable we taken was "Unplanned hospital readmission", in general, if a patient had a readmission in less than 30 days, it would be identified as a case of unplanned hospital readmission. Then we constructed models by Generalized Linear Regression (GLM) and Generalized linear mixed model (GLMM), respectively, for evaluating the potential factors that may be related to hospital readmission. The regression results showed that and we concluded that the patients' medical history, clinical care records and the factor of diabetes were statistically significant related with the risk of readmission.

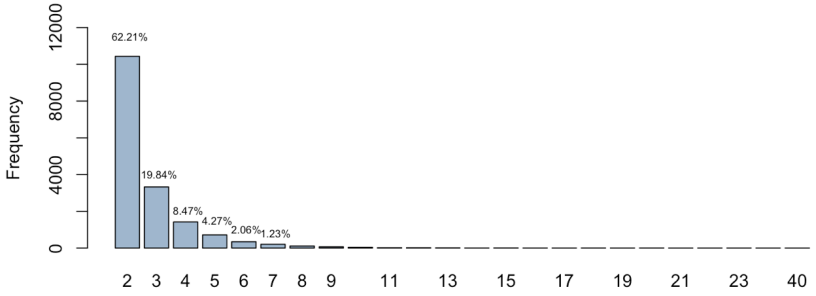
GitHub repo: https://github.com/yangyuanzhe111/A-Population-based-study-of-Hospital-Readmission-Risk-for-American-patients

Methodology

Data

We focused on a data set from the UCI Machine Learning Repository (Data source: https://archive.ics.uci.edu/ml/datasets/diabetes). The dataset originally contains national clinical information data collected from over 130 US hospitals in the period from 1999 to 2008. 71,518 patients had been traced with single or multiple records, 101,766 observations were included with 49 features, such as demographic and geographic information, history medical records, whether a patient was readmitted or not, etc. The index for hospital admission, which was our main interest, was assessed by the number of encounter in this dataset. And we did notice that 16,733 surveyed patients, that is, approximately 23.40% of them eventually had one or multiple experiences of readmission. For these readmitted US patients, we also observed that approximately 62.21% of them were commonly readmitted by twice.

Figure 1. Number of encounters for readmitted patients



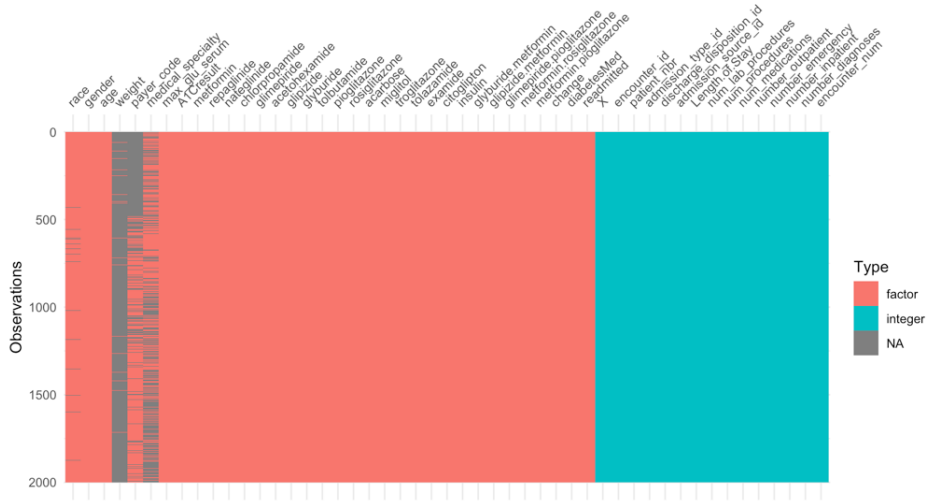
Data cleaning

The process of data cleaning is inevitably necessary for our original dataset because a substantial proportion of missing values has been found in several covariates as shown in Figure 2. We directly removed the variables *weight, prayer_code* and *medical_speciality* since they seem to contain more than 20% unregistered records, even we only randomly sampled 2,000 observations from the pool. Also, the missing values contained in the variable of *race* are removed. Furthermore, since our primary target of this study was exploring the determinants that may associated with the risk of "Unplanned hospital readmission", in general, if a patient had a readmission in less than 30 days, it would be identified as a case of unplanned hospital readmission. Then we created a new binary variable as follows:

$$Y_{ij} = \begin{cases} 1, & if\ patient\ i\ returned\ hospital\ within\ 30\ days\ at\ his\ or\ her\ jth\ encounter \\ 0, & otherwise \end{cases}$$

The cleaning and reformatted dataset eventually contained 69,668 patients with 99,493 observations and 46 variables. For exploring the validation and accuracy of our GLM and GLMM methods in later process, we randomly divided our dataset by a common ratio 80/20 as training and testing datasets, respectively.

Figure 2. Visualization of data structure



Preliminary analysis

Before initiating our modeling, we had a brief overview on our dataset. Table 1 summarized the frequency and relative frequency of the unplanned hospital readmission on categorical variables, such as gender and age groups, the implementation of diabetes medication. Besides, mean and standard deviation were computed for continuous variables, for instance, the inpatient days and number of lab procedures. From Table 1, we found that the odds ratio of being readmitted for female was slightly greater than the one for male patients; the proportion of readmission within 30 days did increase along with older patients. For the category of

ethnicity, no obvious pattern could be intuitively detected. Furthermore, the odds ratio of readmission between diabetic and non-diabetic patients was greater than 1, which was computed by (9.00/67.89)/(2.22/20.89), indicating a likelihood that the diabetic patients would be readmitted to hospital at a higher risk than other patient. Meanwhile, we also noticed that there were differences in the number of inpatients and emergency visits between readmitted patients and non-readmitted patients, implying that the probability of readmission might be infected by these factors. However, we should notice that the dataset had an imbalanced issue since the readmission rate was only 11.22% out of the whole dataset, therefore, the accuracy and efficiency of our later modelization need to be elaborately examined.

Figure 2. Distribution of readmitted rate within demographic variables

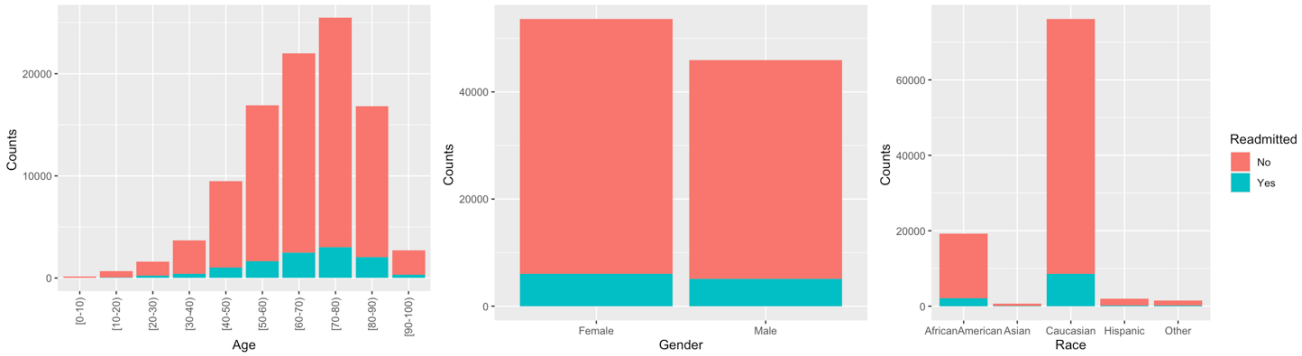


Table 1. Frequency and Relative frequency of the readmission counts

Category	Readmission within 30 days	
	No	Yes
Gender		
Female	47514 (47.76%)	6061 (6.09%)
Male	40809 (41.02%)	5108 (5.13%)
Age		
[0-10]	157 (0.16%)	3 (0.003%)
[10-20]	642 (0.65%)	40 (0.040%)
[20-30]	1379 (1.39%)	232 (0.233%)
[30-40]	3277 (3.29%)	422 (0.424%)
[40-50]	8456 (8.50%)	1009 (1.01%)
[50-60]	15249 (15.33%)	1646 (1.65%)
[60-70]	19519 (19.62%)	2469 (2.48%)
[70-80]	22451 (22.57%)	3018 (3.03%)
[80-90]	14772 (14.85%)	2028 (2.03%)
[90-100]	2422 (2.43%)	302 (0.304%)
Race		
African American	17055 (17.14%)	2155 (2.17%)

Table 1 (Continued). Frequency and Relative frequency of the readmission counts

Asian	576 (0.58%)	65 (0.065%)
Caucasian	67507 (67.85)	8592 (8.64%)
Hispanic	1825 (1.83%)	212 (0.213%)
Other	1361 (1.37%)	145 (0.146%)
Change		
Yes	40447 (40.65%)	5464 (5.49%)
No	47877 (48.12%)	5705 (5.73%)
Diabetes Medication		
No	20783 (20.89%)	2218 (2.22%)
Yes	67541 (67.89%)	8951 (9.00%)
Lab procedures	42.93 ∓19.74	44.21∓19.31
Length of Stays	4.35∓2.98	4.77∓3.03
Number of Medications	15.91∓8.11	16.93∓8.11
Number of Inpatient	0.57 ∓ 1.13	1.23 ∓ 1.96
Number of Emergency	0.18 ∓0.87	0.36 ∓ 1.38

Models

Variable selection

In order to allocate an appropriate combination of explanatory variables for fitting our model, we apply the method of forward and backward stepwise regression proposed by Efroymson [5]. The algorithms within this selection process is straightforward by a series of criteria, such as AIC, BIC and Mallows's Cp. By applying R programming, we have a stepwise selected logistic model with 9 variables: Number of inpatient visits, Diabetes medication prescribed, Number of diagnoses, Number of diagnoses, Metformin, Discharge disposition, Length of stay, Number of medication and Insulin.

Variance Component Model, GLM and GLMM

For detecting the variation in the tendency of being readmitted that could be attributed to the individuals, for a binary response, it has a basic form as follows:

$$Y_{ij} \sim Bernoulli(\pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \mu_i + \varepsilon_{ij}$$

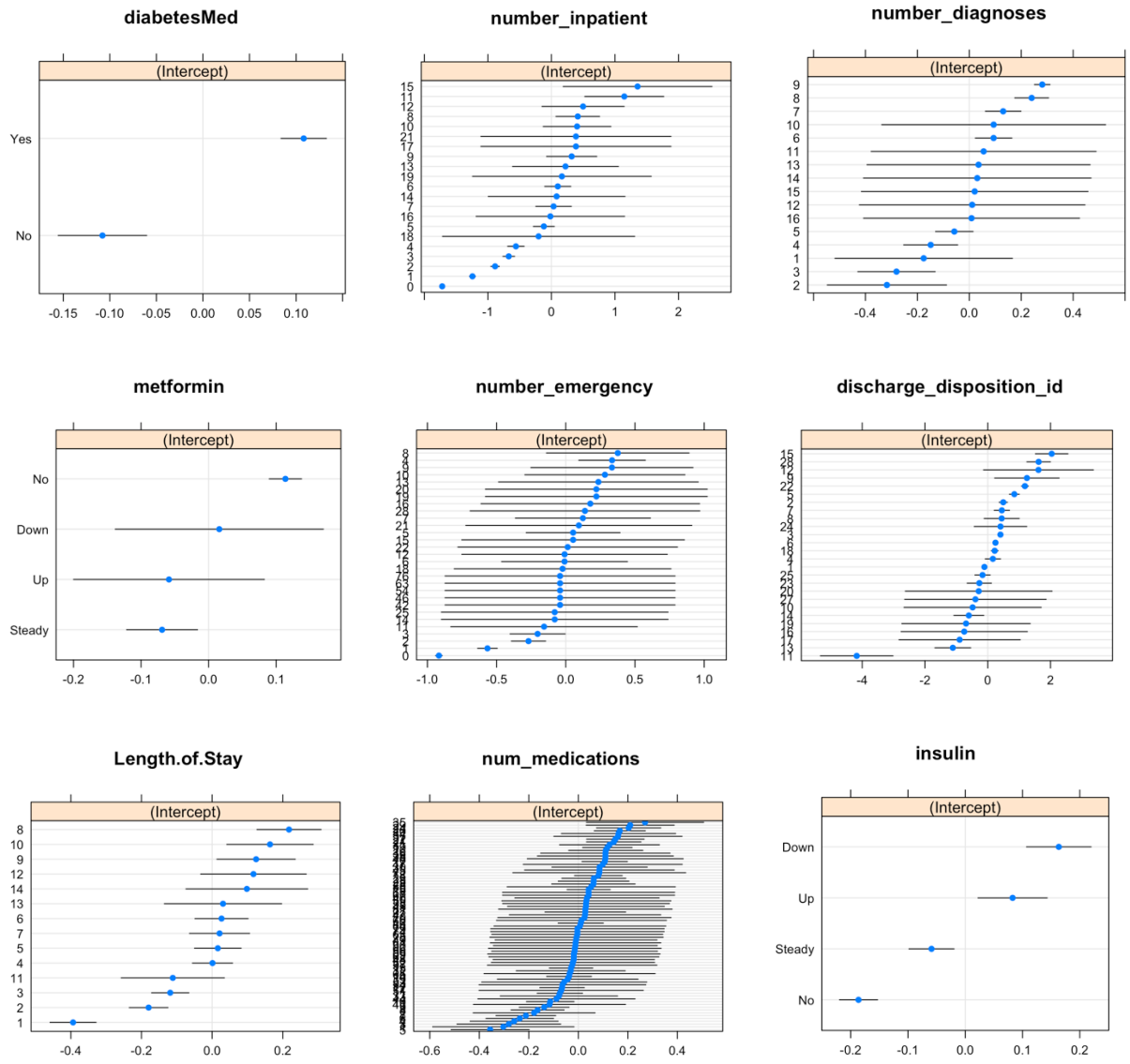
where π_{ij} is the probability of ith individual being readmitted to hospital within 30 days after jth encounter. μ_i it the deviation in the likelihood of readmission for the ith individual from the average. ε_{ij} is the term of error. If a random effect within each individual did exist, an ANOVA test between a null model and a variance component model should have a significant statistic to reject the null hypothesis that the two models have the same performance on fitting data, then random effect across different patients on the likelihood of being readmitted could be confirmed.

Compared to a common generalized linear model (GLM), generalized linear mixed model (GLMM) is an extended form that included both fixed and random effects of covariates. Here, we consider a GLMM containing the individual effect, that is, the average likelihood of readmission was estimated by each random effect within patient i, that is,

$$\text{logit}(\pi_{ij}) = \boldsymbol{X}_i\beta + \mu_i + \varepsilon_{ij}$$

where $\boldsymbol{X}_i$ consists of all the covariates we have selected by the method of stepwise regression above. Theoretically, a GLMM should be more reasonable than GLM, for fitting a longitudinal dataset since it modified the estimation regarding the certain variation within individuals. For verifying this assumption, we executed a ANOVA test between GLM and GLMM, and the accuracy and validation of these two models were confirmed by depicting the ROC curves on testing dataset.

Figure 3. Heterogeneous effects of variables on the probability of readmission



Results

Regression results

We initially conducted an ANOVA test between a null model (only contains intercept term) and a variance model regarding the individual effect. The test statistic is sufficiently large (3491.4) to reject the null hypothesis that there is no difference between these two models. Then we could conclude that the likelihood of readmission did vary across patients. Other than this find, we also noticed that there were variations across different level of variables, for example, the effect of diabetes, number of inpatients, the implementation of insulin did have distinct heterogeneous effects on the probability of readmission, as shown in Figure 3.

Furthermore, the regression result of GLM and GLMM are summarized in Table 2 and Table 3, providing statistically significant estimations for variables: number of inpatients and length of stays, number of diagnoses and emergency visits, as well as the status of diabetes (whether the patient is diabetic or not). Regarding the baseline of our models was patients without being prescribed with diabetic medication during his or her encounter, a diabetic patient definitely has a higher risk of readmission. For instance, if a patient is diabetic, then the odds of being readmitted within in 30 days should be exp(0.197)=1.22 multiply the odds of non-diabetic patient. Likewise, the implementation of insulin could be regarded as an index for evaluating the readmission risk: compared with the baseline, a patient who is prescribed with insulin, his or her odds of readmission is higher than other medicine scenarios. Additionally, patients' medical history and clinical care records, such as the number of inpatients and length of stay, number of emergency he or she had visited could be also affiliated with the classification of patients at risk, considering these records actually refer to physical quality and healthy background.

Table 2. Estimation results by GLM and GLMM regressions

	GLM			GLMM		
	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-2.918	0.168	< 2e-16	-3.111	0.178	< 2e-16
number_inpatient	0.263	0.007	< 2e-16	0.232	0.009	< 2e-16
diabetesMedYes	0.197	0.037	0.000	0.209	0.039	0.000
number_diagnoses	0.050	0.007	0.000	0.052	0.007	0.000
metforminNo	0.013	0.152	0.930	0.013	0.158	0.935
metforminSteady	-0.129	0.153	0.402	-0.135	0.159	0.399
metforminUp	-0.193	0.194	0.320	-0.209	0.201	0.298
number_emergency	0.027	0.009	0.004	0.036	0.011	0.001
discharge_disposition_id	0.025	0.002	< 2e-16	0.025	0.002	< 2e-16
Length.of.Stay	0.016	0.004	0.000	0.015	0.004	0.001
num_medications	0.002	0.002	0.216	0.003	0.002	0.126
insulinNo	-0.113	0.040	0.005	-0.120	0.042	0.005
insulinSteady	-0.113	0.037	0.002	-0.118	0.039	0.002
insulinUp	-0.092	0.044	0.038	-0.104	0.047	0.025

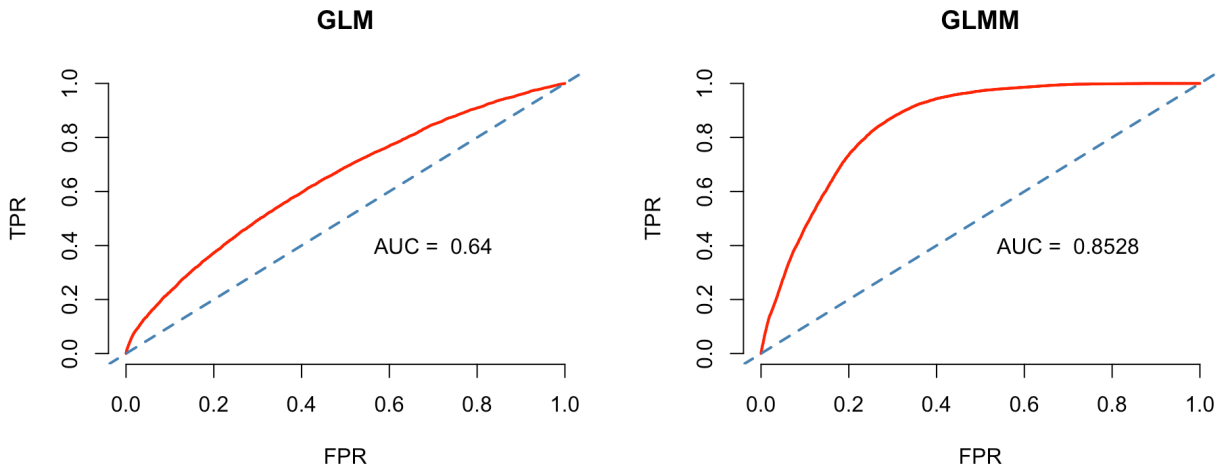
Table 3. MLE's of baseline odds and odds ratios with 95% confidence intervals

	GLM			GLMM		
	Estimate	2.5%	97.5%	Estimate	2.5%	97.5%
(Intercept)	0.045	0.031	0.063	0.054	0.039	0.075
number_inpatient	1.261	1.238	1.284	1.301	1.283	1.320
diabetesMedYes	1.232	1.142	1.329	1.218	1.133	1.310
number_diagnoses	1.054	1.039	1.068	1.051	1.038	1.065
metforminNo	1.013	0.744	1.380	1.013	0.753	1.364
metforminSteady	0.874	0.639	1.195	0.879	0.651	1.188
metforminUp	0.811	0.547	1.203	0.825	0.564	1.206
number_emergency	1.037	1.015	1.060	1.027	1.009	1.046
discharge_disposition_id	1.025	1.021	1.029	1.025	1.021	1.029
Length.of.Stay	1.015	1.007	1.024	1.016	1.008	1.025
num_medications	1.003	0.999	1.006	1.002	0.999	1.005
insulinNo	0.887	0.817	0.964	0.893	0.826	0.966
insulinSteady	0.889	0.824	0.959	0.893	0.831	0.960
insulinUp	0.901	0.823	0.987	0.912	0.836	0.995

Model Validation

For evaluating the validation and accuracy of GLM and GLMM regressions, we predicted the probability of readmission on our testing dataset and depicted the ROC curves as shown in Figure 4. The ROC curves basically display the performance of classification by plotting the true positive rate (TPR) against the false positive rate (FPR). It indicated that the classification by GLMM was substantially better than the one by GLM, since it had a larger value of AUC. Therefore, GLMM model seemed to be more appropriate and efficient for fitting this dataset, and helped us to identify the individual difference (personal effect) in the risk of readmission.

Figure 4. ROC curves between GLM and GLMM on testing dataset



Discussion

In this report, we focused on investigating potential determinants that could be associated with the classification for readmitted patients at risk. We initiated our study by the process of data cleaning and preliminary analysis: both the numerical and graphical evidence suggested that the readmission risk varied across individuals regarding their physical quality and healthy background. For more convinced evidence, we applied the variance component model, GLM and GLMM regressions. The random effect of individuals was proved by ANOVA test, the estimation results by GLM and GLMM indicated that diabetic patients who were prescribed with insulin were suffered with a higher readmission risk than others. Additionally, the number of inpatients and length of stays, number of emergency visits and discharge code could be also regarded as the index for evaluating the readmission risk. The diagnosis of model validation implicated that the GLMM could be more appreciate and accurate for classifying patients refer to readmission.

Although several interesting facts and conclusions have been found and developed, further work need to be considered for improving the integrity of this study. For instance, the initial dataset had an imbalanced issue, although we divided the dataset by training and testing dataset, the potential problem of biased estimations could not be eliminated. More efficient and logical methods should be considered, for instance, K-fold Cross-Validation. Additionally, we selected our variables by the method of stepwise regression on AIC criteria, which is simple and neat, however we may consider to compare it with other methods like LASSO and Elastic net. We also noticed that although the GLMM could be performed as an accurate classification on readmission risk, it was extremely time-consuming by running R, hence we need to improve our computational efficiency and accuracy in our further work.

Reference

[1] Damian M. Predicting Diabetic Readmission Rates: Moving Beyond Hba1c. Curr Trends Biomedical Eng & Biosci. 2017.

[2] Hempstalk, Kathryn & Mordaunt, Dylan. (2016). Improving 30-day readmission risk predictions using machine learning. in Health Informatics New Zealand.

[3] Goudjerkan, Ti'Jay & Jayabalan, Manoj. (2019). Predicting 30-day Hospital Readmission for Diabetes Patients Using Multilayer Perceptron. International Journal of Advanced Computer Science and Applications. 10. 268-275.

[4] Fonarow, G. C., Konstam, M. A., & Yancy, C. W. (2017). The hospital readmission reduction program is associated with fewer readmissions, more deaths: time to reconsider.

[5] Efroymson, M. (1966). Stepwise regression–a backward and forward look. Florham Park, New Jersey.

[6] Myers, R. H., & Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, *29*(3), 274-291.

[7] Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. Journal of statistical Computation and Simulation, 48(3-4), 233-243.

[8] Gandrud, C. (2013). *Reproducible research with R and R studio*. CRC Press.

[9] Zumel, N., Mount, J., & Porzak, J. (2014). *Practical data science with R* (pp. 101-104). Shelter Island, NY: Manning.