How is diabetes readmission rate related to patients' pathologic conditions and medications?

Jiangyue Mao
1003928039

I. Introduction

This project aims to explore the best diabetes readmission rate predictors ("readmitted" variable in the dataset which derives from UCI Machine Learning Repository). The analysis of the variable effect will manifest the rudimentary information of patients, medications and laboratory tests taken during the diabetic encounter to identify patients with worse treatment outcomes and make them targeted to interventions to improve their outcomes and reduce costs by fewer readmission.

II. Methods

1. Variable Selection

First of all, we check the summary of all variables to avoid the tediousness of selection. Since the limited weight data of only 3000 people would cause an error in model building in a size of 20000 training dataset, we exclude the weight variable and omit all other unrecorded data. After checking the summary data of the training dataset, we find out that all patients with acetohexamide, examide, citoglipton, glipizide.metformin, glimepiride.pioglitazone, and metformin.rosiglitazone have the same "No" value (ie. did not take these medications), so these variables are also excluded from the model for no variations. Similarly, since only one or two patients are under the treatment of stolbutamide, troglitazone, and metformin.pioglitazone, the correlations between them and readmitted indicates are small (around $1*10^{-3}$). We perform Chi square test between the readmitted and these variables and the P values are close to 1, indicating the result is not strong enough for us to reject the null hypothesis of independence, hence the exclusion of these variables.

In addition, generalized linear models (GLM) are fitted. We take the first encounter among all the encounters of patients so that each patient is recorded only once and all records are independent. And we fit GLM with all other variables except encounter id, patient nbr, admission source id, and payer code for their absence of actual effects. Secondly, AIC, BIC and LASSO stepwise variable selections are performed, AIC selecting the model with 20 variables, BIC selecting the model with 9 variables. To elaborate on the necessary variables, we perform Lasso selection by the application of cross validation and the selection of 19 variables. Among all the variables that AIC and LASSO select, significant with a p value less than 0.05 are only a few which are employed as predictors to fit new models. Subsequently, we fit generalized linear mixed models (GLMM) with the predictors selected from BIC (model.6) and significant predictors from AIC (model.7) and LASSO (model.8) selections since the large number of predictors from original AIC and LASSO selection lead GLMM function to not converge.

i) Model inferences

Table 1 shows model.3 and 5 with a significantly higher AIC and insignificant Hosmer-Lemeshow test P value, indicating a worse fit. Specifically, we perform ANOVA LRT and Chi square test among all models, and find out that model.1 fits better than all other models. Similarly, ANOVA test among the mixed-effect models reviews model.6 have the lowest AIC and BIC value, with a P value greater than 0.05, indicating a better fit of model.6. However, for their vital role in this dataset to make the best decision, prediction abilities need to be checked in each model.

	AIC	Hosmer-Lemeshow test P value	BIC
readmitted ~ race + gender + age + admission_type_id + discharge_disposition_id + Length.of.Stay + medical_specialty + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + metformin + repaglinide + nateglinide + glipizide + pioglitazone + rosiglitazone + acarbose + diabetesMed (AIC) **Denoted by model.1**	25654	0.1341865	—
readmitted ~ admission_type_id + Length.of.Stay + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed (BIC) **Denoted by model.2**	25900	0.02238069	—
readmitted ~ race + gender + age + admission_type_id + Length.of.Stay + medical_specialty + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + metformin + repaglinide + pioglitazone + acarbose + change + diabetesMed (LASSO) **Denoted by model.3**	26992	$2.658575*10^{-5}$	—
readmitted ~ race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed +Insulin (AIC significant variables) **Denoted by model.4**	25896	0.2438329	—

	AIC	Hosmer-Lemeshow test P value	BIC
readmitted ~ race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed (LASSO significant variables) **Denoted by model.5**	27201	0.003677274	—
lmer(readmitted~admission_type_id + Length.of.Stay + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed+(1| patient_nbr) (BIC) **Denoted by model.6**	35082	—	35172
lmer(readmitted~race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient+number_emergency + number_inpatient + number_diagnoses + diabetesMed + insulin + (1| patient_nbr) (AIC significant variables) **Denoted by model.7**	35082	—	35205
lmer(readmitted~race + gender + admission_type_id + Length.of.Stay + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed + (1| patient_nbr) (LASSO significant variables) **Denoted by model.8**	36952	—	37083

Table 1. AIC and Hosmer-Lemeshow test P value for each model

iii) Model prediction/validation

To explore the model prediction outcomes, we perform cross validation, construct ROC curve, predicted decile plot and calculate mean prediction error (MPE).

First, cross validation is performed to see if the predicted outcomes fit the true values in the test dataset, and we only construct calibration plots for GLMs (Figure 1) for the absence of valid sources of cross validation and calibration plots for GLMMs in my researches. Model.4 has the

smallest deviation of predicted probabilities from the actual probabilities with the smallest mean absolute error.

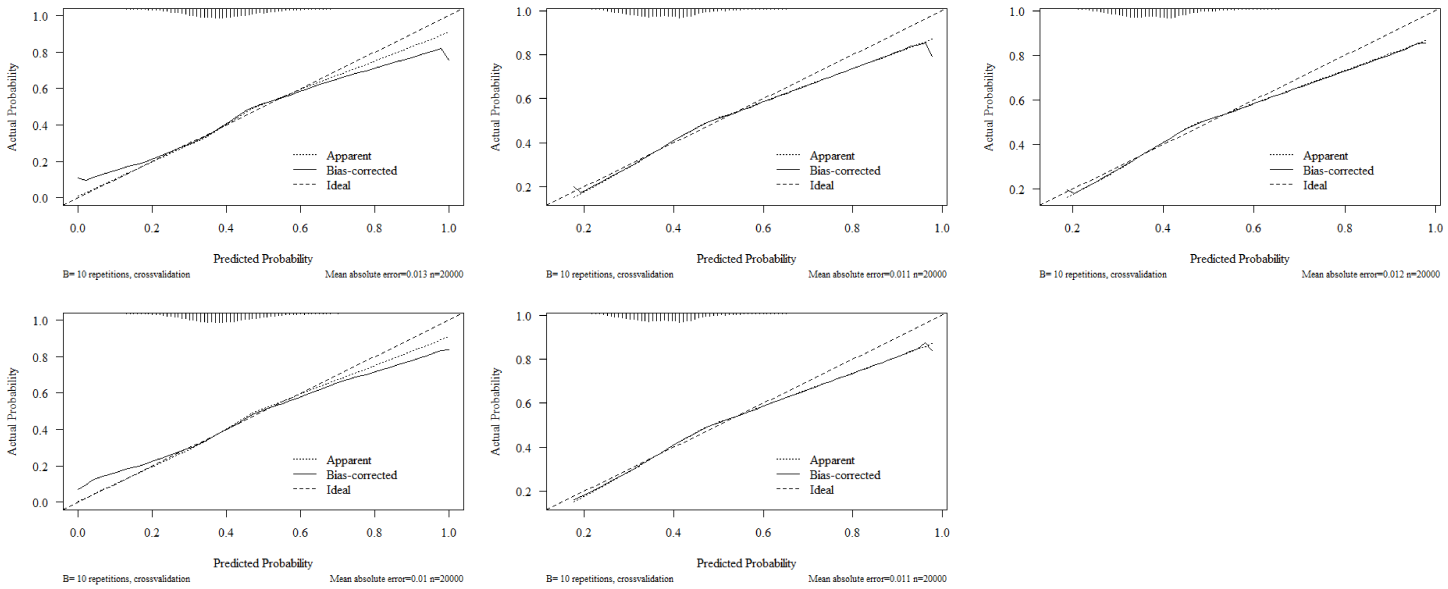

Figure 1. Cross validation for GLM models

Judging from AUC, we find out that mixed-effect models have a better predictability than generalized linear models with larger AUC values. That is, there is 12% greater chance that GLMM models will be able to distinguish between readmitted and non-readmitted people.

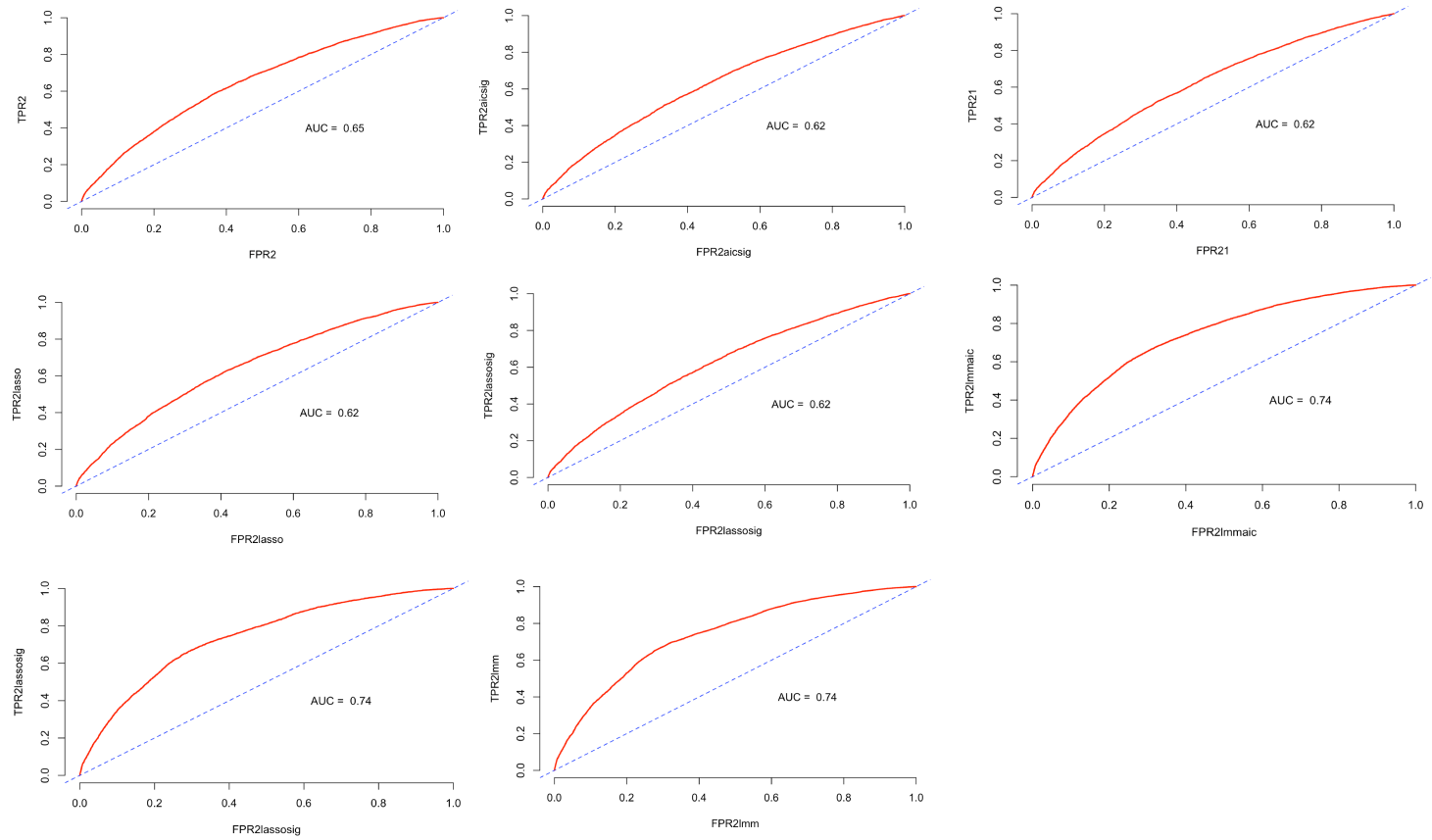

Figure 2. ROC curve and AUC for each model

The decile plots (APPENDIX A) show the differences between predicted probabilities and observed probabilities in the test dataset. Despite the weakness of all models associated with the much larger predicted probabilities than observed probabilities, the mixed-effect models reveal the smallest differences between these two probabilities.

The MPE of all models vary little, with that of model.4 the smallest in APPENDIX B.

2. Results

The above analysis shows that although mixed-effects models have a larger AIC value and fit worse than generalized linear models, the predictability of mixed-effect models is significantly better than that of GLMs. And although they have a larger MPE than GLMs, the difference can be negligible. Moreover, all GLMMs have a similar AUC and a similar difference between predicted probabilities and observed probabilities, with model.6 having the lowest AIC and BIC value among all GLMMs, and we should select the final model as model.6 for its best fit among all GLMMs and its outstanding prediction ability among all models. The variable with the largest coefficient is diabetesMed, being 0.068. To interpret, the odds of being readmitted for people taking diabetes medications is exp(0.068) times that for people not taking diabetes medications controlling for all the other covariates. This means that people not taking diabetes medications have lower odds of being readmitted than the people who take. The ICC of this model is around 0.035, and the 0 in the confidence interval of random effect indicates the lack of variability among the sampled subjects and the insignificant random effects (APPENDIX C). However, since the goal of this project is to find the model with the best prediction ability, we hold model.6 to be the most suitable model and to further perform the diagnostic check.

IV. Diagnostic check of the final model

Diagnostic check for the final model is needed to judge check model assumptions and find outliers in the model. The linear predictor and fitted values vs. deviance residuals plot in Figure.3 shows a similar variation of the linear predictor and fitted values, representing the plot failure to detect any inadequacies in the model. Moreover, the half-normal plot shows closely connected values of individuals, and there is no need to concern about outliers, given the relatively large size of the dataset and the fact that the points are not particularly extreme. We obtain a similar pattern by checking the relationship between residuals and all predictors with predictor vs. residuals plot, and take Length of Stay to exemplify that the residuals are around 0 for all bins to prove that the model fit is acceptable.

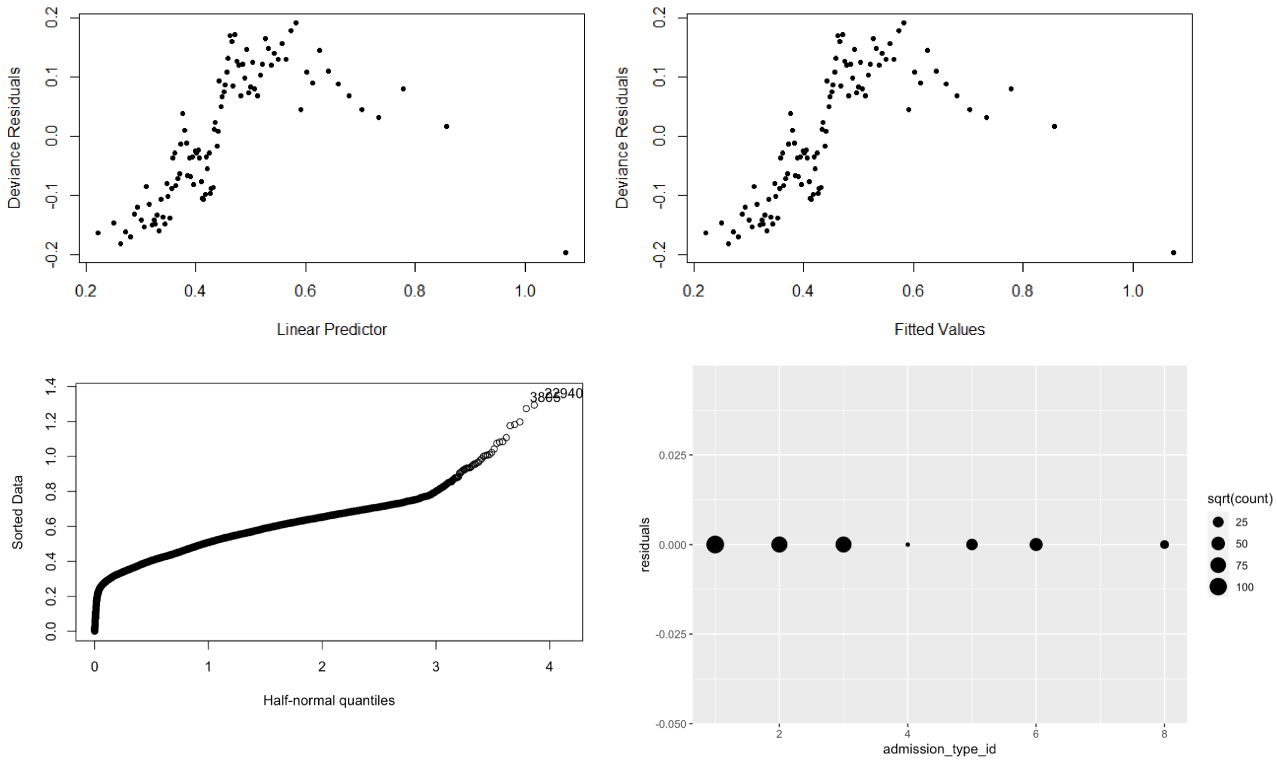

Figure 3. Model diagnostics: linear predictor and fitted values vs. deviance residuals plot, half-normal plot, predictor vs. residuals plot

In conclusion, the final model should be model.6 due to the failure to detect an obvious lack of fit of this model.

V. Description of data

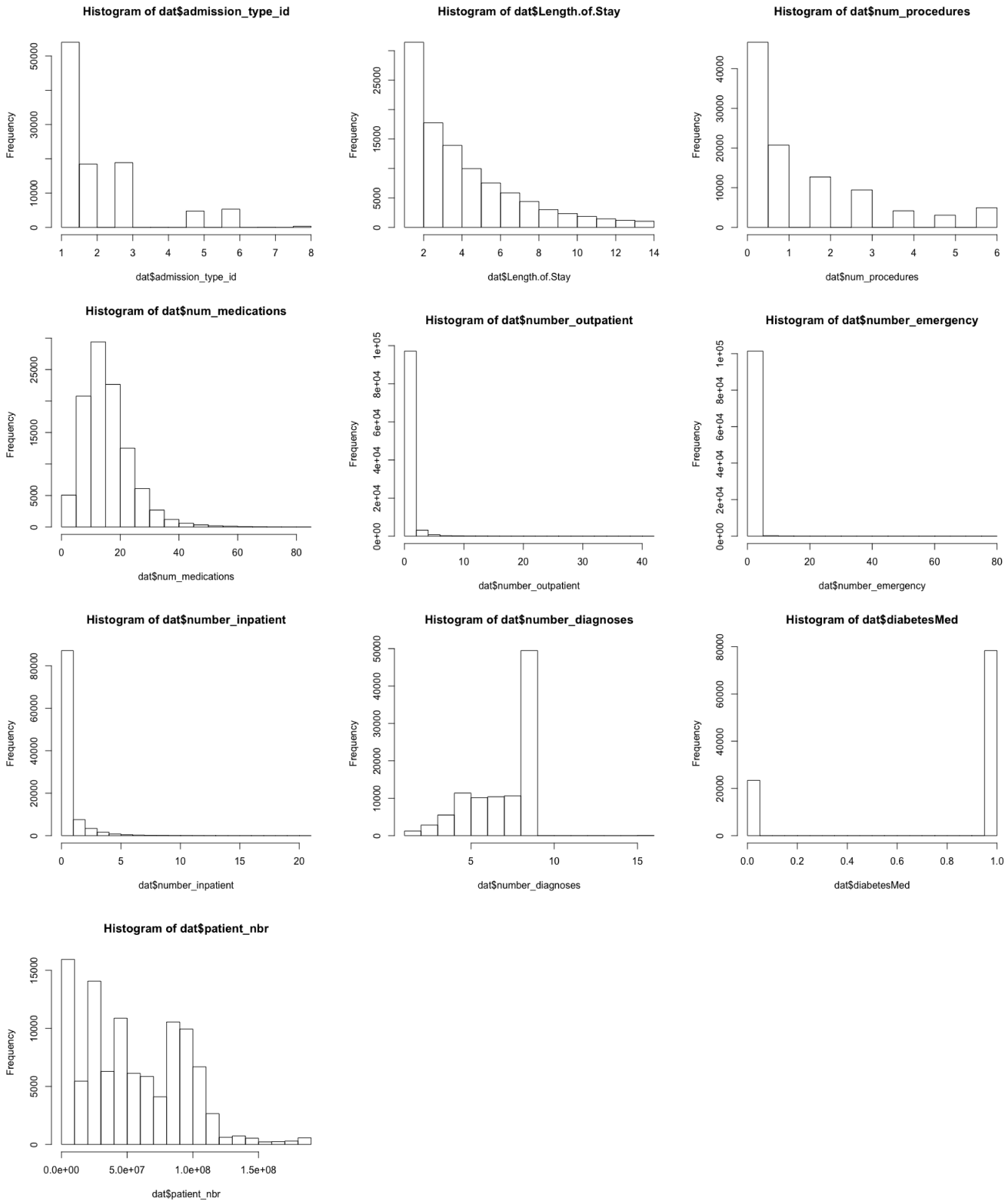

Figure 4. Overview of variables included in the model

Figure 4 shows that most individuals in the sample were admitted as patients in emergency (more than $5*10^4$ people), with relatively short length of stay (more than $3*10^4$ people), a variety of medications (around total number of 10). Most patients were diagnosed nearly 10 times but in a simple treatment procedure during the diabetes encounter.

VI. Scalability of the final model

With test dataset, we check the inferences and predictions of the final model after its selection. The coefficients and variances of fixed as well as random effects are extremely small, suggesting that this model may not have a good fit as shown before. However, the AUC value of this model is 0.77; the predicted probabilities in the decile plot is much closer to the observed probabilities and the MPE is only about 0.208, demonstrating that the predictability of this model is unexceptionable. Therefore, the stability of this model is good in terms of predictability.

VII. Discussion

1. Interpretation and significance of the final model

More length of stay has positive effect on readmission rates since patients staying longer in hospital are usually severe cases which may have more chance to be readmitted. And the history of admission into emergency department are usually associated with readmission since it represents a severe degree of diabetes[1]. In addition, the type and number of medications could also affect the readmission rate since patients taking more various medications may have complications and need multiple treatments[2].

2. Analysis limitations and potential

As discussed before, this model do not have a good fit compared with others, probably due to multicollinearity between variables or large number of levels of some predictors or the incorrect relationship between outcomes and predictors. None of the work to check the multicollinearity, reduce levels and change the relationship to log, exponential and square root produced a model with a significantly better fit. But since we only fit GLM and GLMM in this case, there might be other models that can fit well. Moreover, many patients do not provide information on some variables (NA), so the model might include different predictors compared with the idealized model.The failure of performing calibration plot of cross validation of GLMM might result in the uncertainty of prediction ability of the final model to some extent. However, the good predictability of hospital readmission rate based on AUC, decile plot and MPE might contribute to this model employment in preliminary diabetes studies regarding to readmission rates.

---

[1] Canadian Institute for Health Information, *All-Cause Readmission to Acute Care and Return to the Emergency Department* (Ottawa, Ont.: CIHI, 2012).

[2] Wei, N J et al. "Intensification of diabetes medication and risk for 30-day readmission." *Diabetic medicine : a journal of the British Diabetic Association* vol. 30,2 (2013): e56-62. doi:10.1111/dme.12061

APPENDIX A

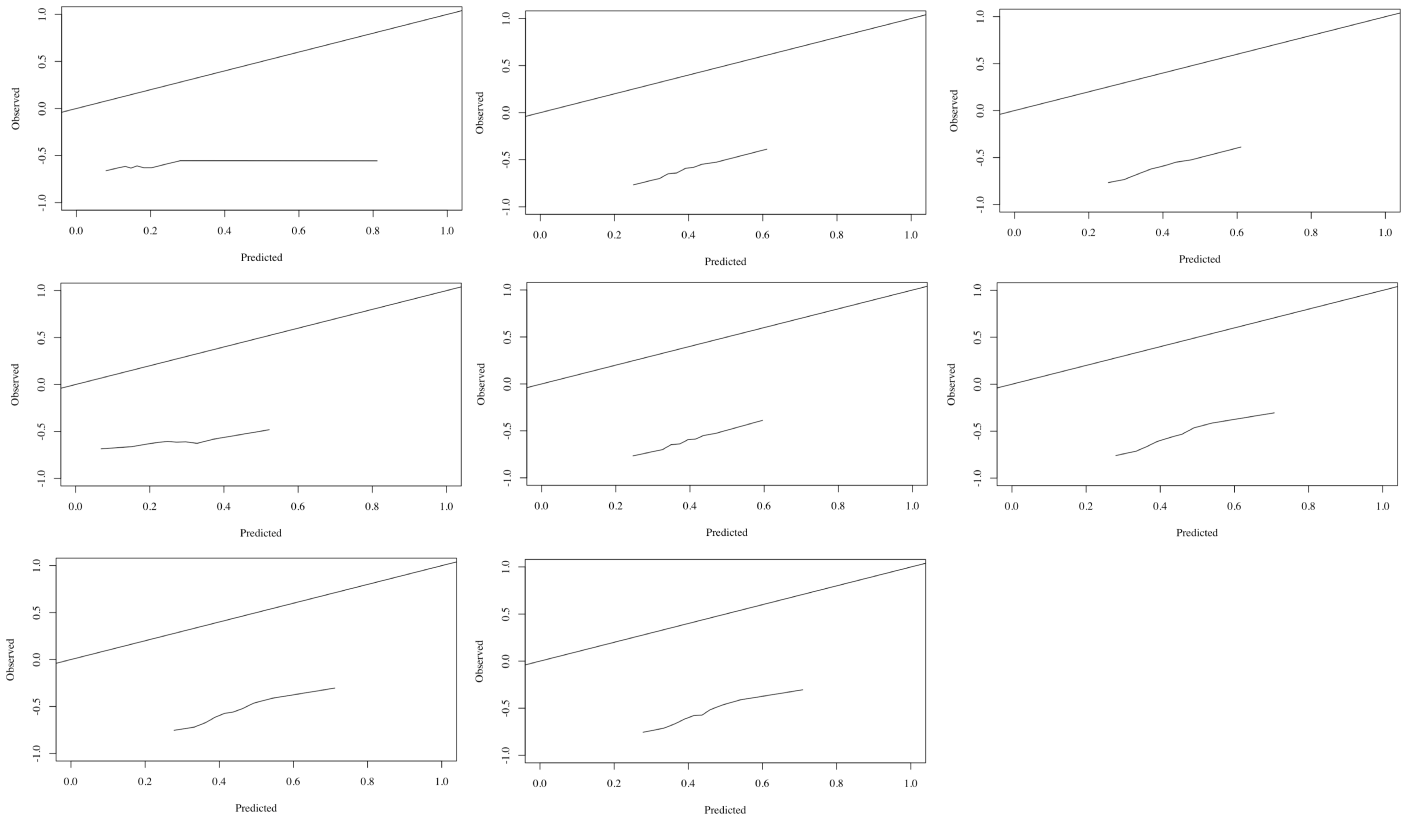

Figure 1. Predicted probabilities vs. observed probabilities for all models

APPENDIX B

	MPE
readmitted ~ race + gender + age + admission_type_id + discharge_disposition_id + Length.of.Stay + medical_specialty + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + metformin + repaglinide + nateglinide + glipizide + pioglitazone + rosiglitazone + acarbose + diabetesMed (AIC) **Denoted by model.1**	0.301612
readmitted ~ admission_type_id + Length.of.Stay + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed (BIC) **Denoted by model.2**	0.2259005
readmitted ~ race + gender + age + admission_type_id + Length.of.Stay + medical_specialty + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + max_glu_serum + metformin + repaglinide + pioglitazone + acarbose + change + diabetesMed (LASSO) **Denoted by model.3**	0.2556627
readmitted ~ race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed +Insulin (AIC significant variables) **Denoted by model.4**	0.2258576
readmitted ~ race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed (LASSO significant variables) **Denoted by model.5**	0.2262724
lmer(readmitted~admission_type_id + Length.of.Stay + num_procedures + num_medications + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed+(1|patient_nbr) (BIC) **Denoted by model.6**	0.2299318
lmer(readmitted~race + gender + admission_type_id + Length.of.Stay+num_procedures + number_outpatient+number_emergency + number_inpatient + number_diagnoses + diabetesMed + insulin + (1|patient_nbr) (AIC significant variables) **Denoted by model.7**	0.2295955
lmer(readmitted~race + gender + admission_type_id + Length.of.Stay + num_procedures + number_outpatient + number_emergency + number_inpatient + number_diagnoses + diabetesMed + (1|patient_nbr) (LASSO significant variables) **Denoted by model.8**	0.229905

Table 1. MPE of all models

APPENDIX C

	Coefficient	Standard error	2.5%	97.5%
(Intercept)	0.1402767	0.0132718	0.113627179	$1.672907*10^{-1}$
admission_type_id	0.0171480	0.0019800	0.013035543	$2.071978*10^{-2}$
Length.of.Stay	0.0058394	0.0011046	0.003492244	$7.987833*10^{-3}$
num_procedures	-0.0094763	0.0018659	-0.013418024	$-5.934244*10^{-3}$
num_medications	-0.0009649	0.0004354	-0.001756901	$-9.827736*10^{-5}$
number_outpatient	0.0268684	0.0031645	0.020741637	$3.345104*10^{-2}$
number_emergency	0.0214645	0.0034590	0.015102313	$2.767311*10^{-2}$
number_inpatient	0.0657403	0.0025404	0.061053653	$7.098574*10^{-2}$
number_diagnoses	0.0223353	0.0015584	0.019181374	$2.548722*10^{-2}$
diabetesMedYes	0.0682618	0.0072724	0.054723963	$8.288745*10^{-2}$
patient_nbr	—	0.0898	0.064150713	$1.107882*10^{-1}$

Table 2. Coefficients, standard error and confidence intervals of all variables in the final model

References

1. Canadian Institute for Health Information, *All-Cause Readmission to Acute Care and Return to the Emergency Department* (Ottawa, Ont.: CIHI, 2012).

2. Wei, N J et al. “Intensification of diabetes medication and risk for 30-day readmission.” *Diabetic medicine : a journal of the British Diabetic Association* vol. 30,2 (2013): e56-62. doi:10.1111/dme.12061.