Are diabetes patients' readmission rate associated with age, HbA1c, length of staying in hospital, medical specialty and other factors? A population-based study of Canada hospital diabetes patients.

STA303 Summer Final Project
Zhang, Ti
ID:1004424517,   2020/08/25
Total word count except table:1479

1. Introduction

Background:
Diabetes is a very prevalent disease that has been linked to numbers of lifestyle factors by many clinical studies. Patients with diabetes tend to have worse medical outcomes than similar patients without diabetes and these patients can be quite costly to the healthcare system. [1] Also, as a diabetes chronic disease, its high readmitted rate of patients can be a heavy medical burden.

Goal of the study:
A large set of data with lots of variables is received, and we need to measure and identify the covariates that have significant impact on the outcome, to find the relations between each covariates and make a model to generalize the result. The goal is to identify patients who are more likely to have worse medical outcome, i.e. have higher readmitted rate, so they can be targeted and interventions can be taken.

The importance of the study:
Obviously, not all diabetes patients presents the same medical burden. The cost of healthcare cannot always be measured directly, but one of the important measures that are commonly collected in health services research that can act as surrogates for both cost and for poor health outcomes is readmission to hospital within a set number of days of discharge from hospital; readmissions are costly and can occur for several reasons, important ones being (a)inadequate care on the initial stay (perhaps after a discharge that was too soon) and (b) generally poor health of the patient, irrespective of the initial length of stay. [1]

Effective interventions can be taken based on the result of this study, and it will bring improvement of treatment outcome and reduction of medical burden.

2. Method selection

Choice of method:
In order to predict the probability of readmission, I choose to use logistic regression without random effect, which is GLM. This requires to make the assumption that each observations are independent, so the original dataset need to be modified as there are duplicate observations for same patient, i.e. the patient. Graphics were used to help in the interpretation of interaction terms in the final model. All of the analysis was performed in R statistical software.

Variable selection:
There are 48 variables in the original file. After sort, regroup and significance test, 10 of the variables are chosen. *Table 1* shows choosing reason and regroup categories, the 10 variables are shaded. Variables will be selected again based on significance.

Table 1. List of variables and their descriptions.

Variable Name	Description	Final group
Race	Original Values: Caucasian, Asian, African American, Hispanic, and others. Regroup into 3 categories which have high significance.	Caucasian
		African-American
		Others
Gender	Original Values: male, female, and unknown/invalid. Removed the unknowns, regroup the rest into 2 categories which have high significance.	Male
		Female
Age	Original Values: [0, 10), [10, 20), . . ., [90, 100). Regroup into 3 categories which have high significance.	<30
		>60
		30-60
Discharge Disposition	Removed the ids representing death or hospice, regroup into 2 categories [2] which have high significance. (According to IDs_mapping.csv provided by UCI [2], we can see that IDs 11,13,14,19,20, 21 represents death or hospice. Observation with these IDs are removed as they cannot be readmitted. The rest of the recordings are regrouped into 2 groups, "*Discharge to home* " and *"Otherwise"*, which have higher significance.)	Discharge to home
		Otherwise
Length of Stay	Regroup into 3 categories which have high significance.	1-5
		6-10
		>10
Medical Specialty	Regroup into 6 categories based on department [3] which have higher significance. Although the original data has 49% missing, it still has a high significance on the model prediction. Medical specialty categories are grouped based on [4] to simplify analysis process.	Other
		Internal-Medicine
		Emergency
		General
		Surgery
		Missing
Number of Procedures	New variable. Represents number of lab tests performed. Base on:"num_lab_procedures","num_procedures","num_medications"	
Number of Visits	New variable. Represents number of visits of patients. Keep as the new variable has significant effect. Based on "number_outpatient","number_emergency", number_inpatient".	0-4
		5-9
Number of Diagnoses	Number of diagnoses entered to the system. Keep as its p-value significant	
HbA1c	New variable. Keep as the new variable has significant effect. Combined variable *"A1Cresult"* and *"change",* regrouped into 4 categories which have higher significance. *"No test was performed"* represents the observations with *"A1Cresult"* stating *"None"*; *"Normal result of the test"* represents the observation with *"A1Cresult"* stating *"Norm"* or *">7"*; "*Result was high but the diabetic medication was not changed"* counts the observations that have *"A1Cresult"* stating *">8"* and *"change"* stating *"No"*; *"Result was high and the diabetic medication was changed"* counts the observations that have *"A1Cresult"* stating *">8"* and *"change"* stating "*Yes"*.	No test performed
		Normal result
		High, no change*
		High, changed*
Number of	These 2 variables are based on data of the 24 featured medications. That indicates whether the drug was	

Medication Taken	prescribed or there was a change in the dosage. Values: *"up"* if the dosage was increased during the encounter, *"down"* if the dosage was decreased, *"steady"* if the dosage did not change, and *"no"* if the drug was not prescribed [5]. *"Number of Medication Change"* counts the values except *"steady"* and *"no"*, *"Number of Medication Taken"* counts the number of values except *"no"*	
Number of Medication Change		
Max Glucose Serum Test Result	Keep as it's an important method to measure diabetes patients' condition	>200
		>300
		None
		Norm
Encounter ID	Removed as it's an identification variable	
Patient Number	Removed after selected test and train dataset, as it's also an identification variable when using GLM (duplicate data for same patient was removed).	
Weight	Removed as it has too many N/A	
Admission Type Id	Removed as p-value not significant	
Diabetes Medication		
24 Featured Medications	Including: metformin, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, -roglitazone, tolazamide, examide, citoglipton, insulin, metformin-pioglitazone, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone. Regrouped into " **Number of Medication Taken**" and " **Number of Medication Change** "	

Model Violations and Diagnostics:

The independence of each observations is assumed for GLM model, so we need to drop duplicate data, i.e. keep only one record for each patient based on the variable patient number. Also, the response variable, *readmitted rate,* has 3 categories. It has been reorganized into a binary data. The reorganization rule is:

(1) Consider as readmission if readmitted in 30 days.
(2) Consider as no-readmission if readmitted in more than 30 days/ no readmission.

In order to diagnose the model properly, following measures are taken:

(1) Check if there is any missing value
(2) Do the residual check to see if:
    a) The variables are independent.
    b) The result follows normal distribution.
(3) Predict the model accuracy: Divide the dataset into test and train parts. Create a test dataset which contains 20000 random selection of patients. Then use train dataset to fit and predict the data in test dataset. Finally compare with the true value in test dataset to see if accurate.
(4) Operating the ROC curve to see if the model has a good discrimination ability through AUC.
(5) Compare the AIC value to see which model (before reduction vs. after reduction) is better.
(6) Perform an internal validation using cross-validation to see if the model perform well in predicting the response from the training dataset.

3. Result section

Description of data

Numerical and visual summary express in Table2, 3, 4.

Table 2 shows the distribution of variable values and readmission under 99342 observations. *Table 3* expressed the coefficients of non-interaction terms estimated from the final logisitic regression model. *Table 4* focus on the correlation between readmission and other covariates. The final model is sorted and interpreted into *Table 3* and *Table 4.*

According to *Table 2*, measurement of HbA1c was infrequent, occurring in only *16.94%** of encounters where diabetes medication was included. Among those patients whose test result was recorded, *48.35%***of them were less than 8%. Among those who have test result >8%, *65.17%**** have medication change documented.

From *Figure 1* we can get a more intuitive understanding of predicted readmission rate with respect to HbA1c. For better understanding purpose, more interpretation of final data and model will be shown in part " ***Final Model Interpretation***".

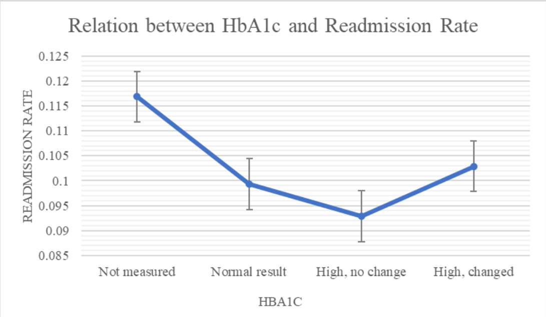


FIGURE 1

**16.94% = 8.75%+2.85%+5.34%*

***48.35% = (2.85%+5.34%)/16.94%*

****65.17% = 5.8%/(5.8%+3.1%)*

Table 2. Distribution of variable values and readmission(population size 99342). Naming method partial referenced [5]

Variable	Num of observe	% of study population	Readmitted	
			Num of observe	% of study population
HbA1c				
No test was performed	82508	83.05%	9641	11.68%
Normal result of the test	8697	8.75%	864	9.93%
Result was high but the diabetic medication was not changed	2831	2.85%	263	9.29%
Result was high and the diabetic medication was changed	5306	5.34%	546	10.29%
Race				
African American	18772	18.90%	2149	11.45%
Caucasian	74222	74.71%	8556	11.53%
Other	6348	6.39%	609	9.59%
Gender				
Female	53455	53.81%	6128	11.50%
Male	45887	46.19%	5186	11.30%
Age				
<30	2499	2.52%	279	11.20%
>60	66412	66.85%	7920	11.90%
30-60	30431	30.63%	3115	10.20%

Discharge Disposition				
Discharge to home	60232	60.63%	5602	9.30%
Otherwise	39110	39.37%	5712	14.60%
Length of Stay				
>10	5300	5.34%	668	12.60%
1~5	71579	72.05%	7584	10.60%
6~10	22463	22.61%	3062	13.60%
Medical Specialty				
Other	4330	4.36%	379	8.76%
InternalMedicine	23899	24.06%	2684	11.20%
Emergency	7420	7.47%	845	11.40%
General	7252	7.30%	879	12.10%
Surgery	7826	7.88%	771	9.86%
Missing	48615	48.94%	5753	11.80%
Number of Visits				
0~4	93027	93.64%	9816	10.60%
5~9	6315	6.36%	1498	23.70%
Max Glucose Serum Test Result				
>200	1419	1.43%	184	13.00%
>300	1188	1.20%	179	15.10%
None	94189	94.81%	10657	11.30%
Norm	2546	2.56%	294	11.50%

Process of obtaining the final result

After having the original resorted and recategorized, we removed the duplicate observations for same patients, leaving each patient one observation to fit the GLM model. This dataset will be called "GLM dataset" for convenience from now on. Some covariant with less significance are also removed to simplify the model. This process reduced the dataset size and gives us a new dataset with each observation independent.

In order to test the accuracy of prediction of the model, the GLM dataset was separated into 2 parts, including training and testing. The test dataset contains 20000 random selection of patients, which will never be used for modeling . Then use train dataset to fit and predict the data in test dataset. Finally compare with the true value in test dataset to see if accurate. Also ROC curve is constructed to see if the model has a good discrimination ability.

After receiving the logistic model produced from GLM dataset, we can see that there are significant variables, but still some variables with low significance existing. Therefore we need to reduce the model by using *step()* function. It gives us a reduced model with unnecessary variables removed. From this reduced model we can see which variable has high impact on readmission rate, and we can move on to find the correlation between each of them. Result of this part in *Table 3 and Table 4.*

Goodness of Final Model

The final model validation can be seen in *Table 3* and *Table 4*. Further explanation and assumption verification are in ***Final Model Interpretation*** for better understanding purpose.

Table 3. Coefficients of non-interaction terms estimated from the final logistic regression model. Naming method partial referenced [5]

		Estimate	P value
	Intercept*	-4.0790	1.12E-04
Medical Specialty	Other	**Reference**	
	InternalMedicine	0.1829	0.0471 *
	Emergency	0.0228	0.8373
	General	0.1951	0.0591 .
	Surgery	-0.0645	0.5349
	Missing	0.1810	0.0432 *
	Number of Diagnoses	0.0600	<0.0001***
	Number of Medication Changes	0.0990	0.0055 **
	Number of Medication Taken	0.0356	0.0490 *
	Number of Procedures	0.0013	0.0856
Discharge Disposition	Discharge to home	**Reference**	
	Otherwise	0.5088	<0.0001***
Length of Stay	>10	**Reference**	
	1~5	0.0546	0.4496
	6~10	0.2007	0.0054 ***
Number of Visits	0~4	**Reference**	
	5~9	0.7061	<0.0001***
HbA1c	No test was performed*	**Reference**	
	Normal result of the test*	-0.0772	0.1691
	High result, medication not changed.*	-0.0939	0.3527
	High result, medication changed.*	-0.0584	0.4194
Age	<30	**Reference**	
	>60	0.1166	0.3390
	30-60	-0.0794	0.5171
Race	AfricanAmerican	**Reference**	
	Caucasian	0.0126	0.7724
	Other	-0.0491	0.5068
Gender	Female	**Reference**	
	Male	0.0317	0.3210
Max Glucose Serum Test Result	>200	**Reference**	
	>300	-0.1052	0.5961
	None	0.0333	0.8015
	Norm	0.0881	0.5737

*Coefficients significant at the 0.001 significance level.

**Variable HbA1c has 4 categories in measurements: "Not measured" acting as a reference; "Normal" represents normal result of the tes;, " High, no change" represents result was high but the diabetic medication was not changed; "High, changed" represents result was high and the diabetic medication was changed.*

Table 4. Coefficients of interaction terms estimated from the final logisitic regression model. Naming method partial referenced [5]

Attribute	Value	Attribute	Value	Estimate	P value	
Age	>60	Medical Specialty	InternalMedicine	-0.8227	0.0744	.
			Emergency	-1.4351	0.0139	*
			General	0.2227	0.7794	
			Surgery	9.8143	0.8989	
			Missing	-1.4014	0.0001	*
	30-60		InternalMedicine	-0.7531	0.1114	
			Emergency	-1.3906	0.0207	
			General	0.2156	0.7887	
			Surgery	9.9148	0.8979	
			Missing	-1.2520	0.0010	*
Number of Diagnoses		Discharge Disposition	Otherwise	-0.0602	0.0004	***
Race	Caucasian	Discharge Disposition	Otherwise	0.0160	0.8504	
	Other			0.3737	0.0111	*
Medical Specialty	InternalMedicine	Discharge Disposition	Otherwise	-0.4698	0.0093	**
	Emergency			-0.2216	0.3086	
	General			-0.4748	0.0195	
	Surgery			-0.1559	0.4524	
	Missing			-0.5134	0.0033	**
	InternalMedicine	Age	>60	-0.8227	0.0744	.
			30-60	-0.7531	0.1114	
	Emergency		>60	-1.4351	0.0139	*
			30-60	-1.3906	0.0207	*
	General		>60	0.2227	0.7794	
			30-60	0.2156	0.7887	
	Surgery		>60	9.8143	0.8989	
			30-60	9.9148	0.8979	
	Missing		>60	-1.4014	0.0001	***
			30-60	-1.2520	0.0010	**
HbA1c	Normal result	Number of Diagnoses		0.0219	0.4820	
	High, no change			0.1134	0.0151	*
	High, changed			-0.0194	0.5624	

*Coefficients significant at the 0.001 significance level.

4. Discussion Section

Final Model Interpretation

Without considering other coefficients:

With respect to readmission and without considering other coefficients, measurement of HbA1c has not shown a very significant association with the readmission reduction rate, which can be found in *Table 3*: *P(Normal*)=0.1691>0.05, P(High, no change*)=0.3527>0.05, P (High, changed*)=0.4194>0.05*.

Among the other covariates, Medical Specialty in Internal Medicine *(P=0.0471)*, in Missing *(P=0.0432)*; Number of Diagnoses *(P<0.0001)*; Length of Stay in 6~10 days *(P=0.0054)*; Discharge Disposition to Otherwise *(P<0.0001)*; Number of Medication Changes *(P=0.0055)*; Number of Medication Taken *(P=0.0490)* and Number of Visits in 5~9 times *(P<0.0001)* show high significance with respect to readmission. Detailed information can be seen from *Table 3*.

Since the Max Glucose Serum Test Result variable *(P=0.5961, P=0.8015, P=0.5737)* and Gender *(P=0.3210)* variable are not significant in the model, they are removed from future analysis. Detailed data distribution can be seen from *Table 3*.

With considering other coefficients:

We then examined the co-relationship between readmission and HbA1c adjusting for covariates such as Number of Medication Taken and age. When using the logistic model testing the covariant interaction, the result of coefficients shows a exponential growth. So again, by using function step(), we obtained a reduced model with much fewer coefficients. Some of the interaction terms gives significant p-value (*Table 4*). This suggests a relationship between thesevariables.

The significant pairwise interactions between the covariates are: number of diagnoses with discharge to other than home *(P=0.0004)*; medical specialty in internal medicine with discharge to other than home *(P=0.0093);* medical specialty in general with discharge to other than home *(P=0.0195)*; medical specialty in emergency with age>60 *(P=0.0139)*; medical specialty in emergency with age30-60 *(P=0.0207)*; Race other than African American and Caucasian with discharge to other than home *(P=0.0111)*; HbA1c test result high and no medication change with number of diagnoses *(P=0.0151)*

There was no significant interaction among the other coefficients. Only these interactions were included in the final model.

The final model suggest that the relation of readmission probability and the number of diagnoses significantly depends on the disposition patients discharged to. Also, HbA1c has a significantly effect on readmission rate. Predicted probability of readmission with respect to HbA1c shown in Figure 1.

Limitations of Analysis

The biggest limitation of this analysis is that we choose a GLM model, which requires the assumption of independence. Due to this reason a lot of data was removed from out final dataset. This significantly reduce the size of total observation and may cause inaccuracy of analyzation. Also , this analysis is designed based on high conservation criteria. A number of observations are removed due to incomplete recording or N/A existing, and some regrouping might not be the best choice. This is certainly an underestimate of the actual circumstance.

Actual relevance

This model can help hospital and other organization take better tretments to cure diabetes patients according to their relevancy with the readmission rate, this is certainly a good way to reduce the readmission rate and therefore the medica burden can be effectively reduced.

5. Appendix

Figure 2

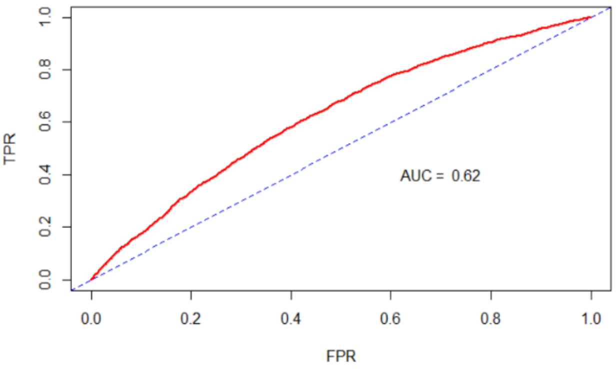


Figure 3

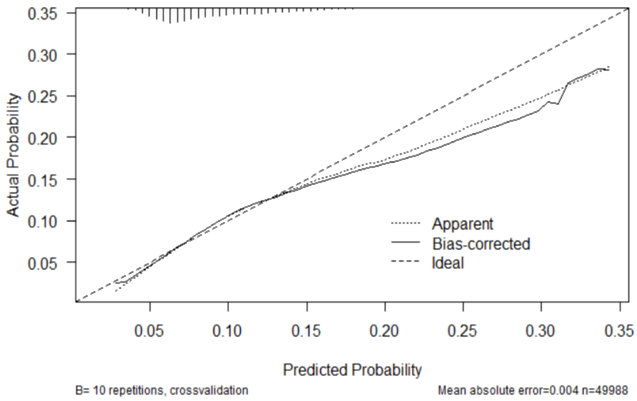
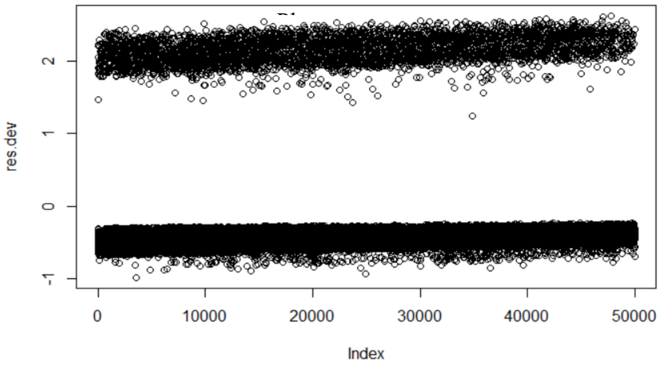


Figure 4 Residual



6. References

[1] M. K. A. Khan, "STA303H1S/STA1002HS Final Project," 2020.

[2] A. Long, "Using Machine Learning to Predict Hospital Readmission for Patients with Diabetes with Scikit-Learn," *Towards Data Science,* 21 October 2018.

[3] kernel1e0f8b3420, "Diabetes 130 US Hospital With GBM and Xgboost - for years 1999-2008," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/tim101/diabetes-130-us-hospital-with-gbm-and-xgboost.

[4] S. G. University, "The Ultimate List of Medical Specialties and Subspecialties," *The SGU Pulse-Medical School Blog,* pp. https://www.sgu.edu/blog/medical/ultimate-list-of-medical-specialties/, 11 12 2017.

[5] 1. J. P. D. C. G. J. L. O. S. V. Beata Strack, "Impact of HbA1c Measurement on Hospital Readmission Rates:," *BioMed Research International, Hindawi Publishing Corporation,* Vols. Volume 2014, Article ID 781670, 11 pages, 2014.