QMST 5334 Mid-term Report

Prediction of hospital readmissions for patients with diabetes

Neeraj Kumar Reddy Panta, Ruchi Dilip Kukde

The prediction of hospital readmission is a significant healthcare research area from data analytics and information systems perspective. It aims to develop and analyze models using historical medical data to predict probability of a patient returning to hospital in a certain period, e.g., 30 or 90 days, after the discharge (Wang & Zhu, 2022). Prediction of hospital readmission is a complex research problem due to the intricate nature of various diseases and healthcare eco-systems (Hospital Readmissions, 2018). As data scientists, we can provide solutions to the healthcare sector for optimizing resources and reducing the readmissions and associated costs using technology and various analytical tools in hand. Apart from the tangible outcomes associated with these solutions, the development and implementation of analytical models for hospital readmissions is significant from humanitarian point of view: in a way that helps patients with better treatment, care, and support (Healthstream, 2021). The motivation of choosing this study is to apply data analytics, specifically predictive analytical tools to identify underlying causes for readmissions, to achieve meaningful and transparent predictions for effective decision making. The report is divided into five sections describing purpose and description of dataset, basic statistics, descriptive plots, estimation of parameter values and test of significance.

A. Purpose and description of dataset

The dataset used for this project focuses on hospital readmissions data in United States. The data was collected in the form of comprehensive clinical records across hospitals throughout United States by Health Facts database – Cerner Corporation, Kansas City. The data was submitted to UCI Machine Learning Repository (UCI Machine Learning Repository, 2014) in 2014 on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data (Strack, et al., 2014). The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 55 features representing patient and hospital outcomes. Appendix A on pages 6 and 7 of the report presents the data dictionary for the variables in the dataset. Information was extracted from the database for encounters that satisfied the following criteria:

i. It is an inpatient encounter (a hospital admission).
ii. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
iii. The length of stay was at least 1 day and at most 14 days.
iv. Laboratory tests were performed during the encounter.
v. Medications were administered during the encounter.

B. Basic Statistics

Exploratory analysis is the first and critical step in data analysis process. It involves performing initial investigations on data in order to discover patterns, spot anomalies, test hypothesis and to check assumptions. It is generally carried out by description of distributions through summary statistics and graphical representations. For this study, we have chosen variable of significance from the dataset as Time in Hospital. It is also called Length of Stay (LOS). It is a clinical metric that measures the length of time elapsed between a patient's hospital admittance and discharge. Figure 1 shows an illustration of the dataset indicating the data variable 'time_in_hospital'.

	A	B	C	D	E	F	G	H	I	J
1	encounter	patient_nl	race	gender	age	weight	admission	discharge_	admission	time_in_hospital
2	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1
3	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3
4	64410	86047875	AfricanAm	Female	[20-30)	?	1	1	7	2
5	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2
6	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1
7	35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3
8	55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4
9	63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5
10	12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13
11	15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12
12	28236	89869032	AfricanAm	Female	[40-50)	?	1	1	7	9
13	36900	77391171	AfricanAm	Male	[60-70)	?	2	1	4	7
14	40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7
15	42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10

Figure 1: Illustration of 'time_in_hospital' variable from dataset(UCI Machine Learning Repository, 2014)

In order to visualize the basic statistics, we noted that the data type of the variable as numeric, (specifically integer) with the units as 'days'. It was observed that no null values exist for the variable and the exploratory analysis can be carried out on 101766 observations. Table 1 shows the measures of centre and spread obtained for length of stay. It is observed that the average length of stay is 4.39 days. The national average for a hospital stay is 4.5 days, according to the Agency for Healthcare Research and Quality, at an average cost of $10,400 per day. The statistical analysis of this variable is significant as it is an important indicator of efficiency of hospital management, patient quality of care, functional evaluation. It is reported that

shorter hospital stays reduce the burden of medical fees, increase the bed turnover rate. This in turn increases the profit margin of hospitals, while lowering the social costs. Additionally, it is important for further analysis: determination of impact of length of stay on readmission risk.

Table 1: Summary of basic statistics

Measures of Location	
Mean	4.395987
Median	4
Mode	3
Measures of spread	
Range	13
IQR	4
Variance	8.910868
Standard Deviation	2.985108

C. Descriptive plots (histograms and box plots)

Figure 2 shows the graphical description of the variable in the form of histogram (including mean, median and mode) and boxplots. The plots depict that the variable is a skewed distribution. The box plots show that the minimum and maximum values are 1 and 14 respectively while the three quartiles are 2, 4 and 6 respectively.

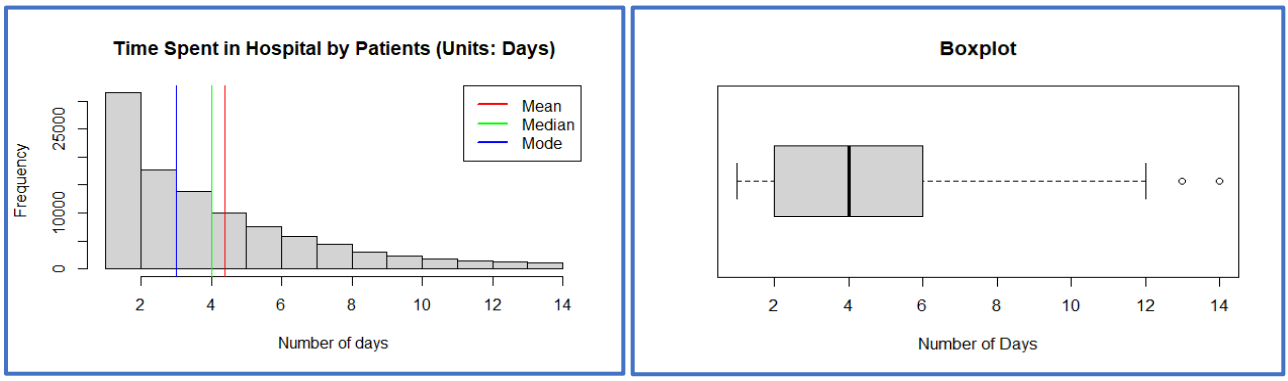

Figure 2: Graphical description of time in hospital using histogram (left) and boxplot (right)

The findings of the basic statistical analysis are: (i) Mean of time in hospital (4.395 days) obtained for this dataset is close to the national average of 4.5 days. (ii) Descriptive statistics indicate that time in hospital is a positively skewed distribution. (iii) 50% of the patients spent 2-6 days in the hospital (iv) 2252 patients who stayed in hospital for 13 days and 14 days are identified as outliers. Consequently, it was intriguing to understand that time in hospital could be analysed in association with other variables in the dataset from different perspectives. It would be interesting to determine correlation of length of stay with readmissions and severity and type of diagnosis variables in the dataset: whether patients with a lower length of stays were readmitted more frequently to the hospital, what were the diseases for such patients, how many medications were administered, were any lab procedures carried out?

D. Estimation of parameter values

With an objective to estimate the population mean μ, we added two factors to the variable of significance. The two factors namely ‘readmitted’ and ‘gender’ were used for this analysis. Figure 3 shows the illustrated example of a part of dataset depicting these factors. The factor variable ‘readmitted’ is a categorical variable which signifies whether a patient was readmitted in less than 30 days, greater than 30 days or not reported to be readmitted. Gender represents male and female patients in the dataset. It was noted that both factors are categorical variables with no null instances. In order to perform estimation, the dataset was categorized into six groups.

	A	B	C	D	E	F	G	H	I	J	AX
1	encounter	patient_nl	race	gender	age	weight	admission	discharge_	admission	time_in_h	readmitted
2	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	NO
3	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	>30
4	64410	86047875	AfricanAm	Female	[20-30)	?	1	1	7	2	NO
5	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	NO
6	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	NO
7	35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3	>30
8	55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4	NO
9	63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5	>30
10	12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	NO
11	15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	NO
12	28236	89869032	AfricanAm	Female	[40-50)	?	1	1	7	9	>30
13	36900	77391171	AfricanAm	Male	[60-70)	?	2	1	4	7	<30
14	40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	<30
15	42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	NO
16	62256	49726791	AfricanAm	Female	[60-70)	?	3	1	2	1	>30
17	73578	86328819	AfricanAm	Male	[60-70)	?	1	3	7	12	NO
18	77076	92519352	AfricanAm	Male	[50-60)	?	1	1	7	4	<30
19	84222	1.09E+08	Caucasian	Female	[50-60)	?	1	1	7	3	NO

Figure 3: Illustration of factors 'readmitted' and 'gender' from dataset (UCI Machine Learning Repository, 2014)

The six groups are G1: Female patients readmitted within (less than) 30 days, G2: Female patients readmitted after 30 days, G3: Female patients who were not reported as readmitted, G4: Male patients readmitted within (less than) 30 days, G5: Male patients readmitted after 30 days and G6: Male patients who were not reported as readmitted. The histograms for the above-mentioned groups are shown in Figure 4. All the groups exhibit right side / positive skewness. The comparison of the distributions is also seen using box plot representation as shown in Figure 5.

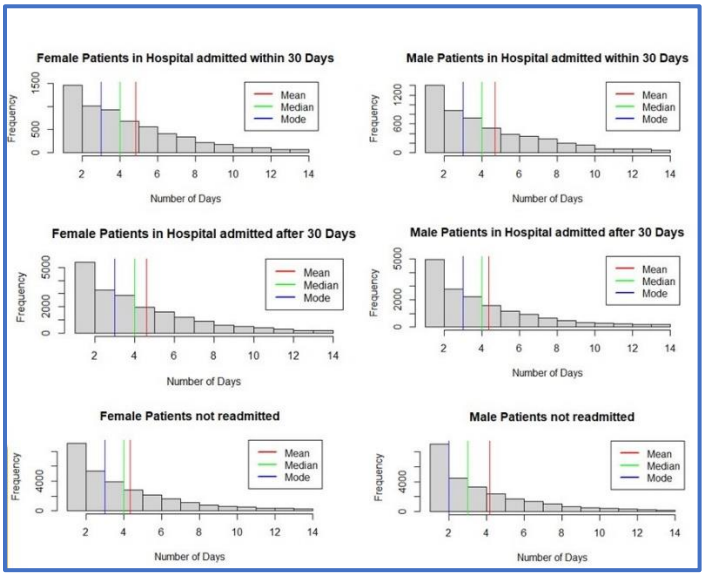

Figure 4: Graphical description of the six groups using histogram plots

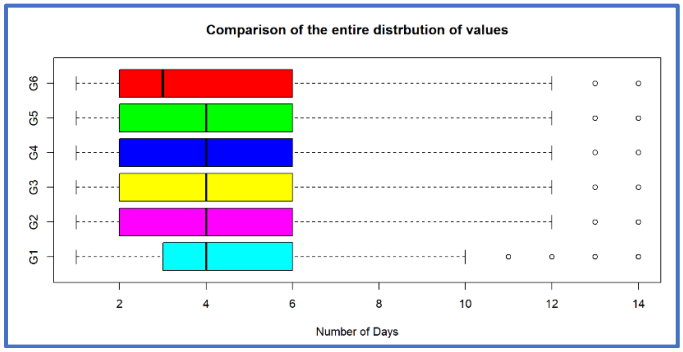

Figure 5: Graphical description of the six groups using comparative box plot visualization

The sample mean values also referred as point estimates are computed for the six categories. These values are summarized in Table 2. If we recall the average mean of length of stay for the entire dataset, it was calculated as 4.39 days. We can clearly observe that the sample estimates for μ vary across the different sub-samples. The 95 percent confidence intervals were also computed for the six groups. The 95 percent confidence bands for each of the groups obtained through the t.test function are tabulated in Table 3. From Table 3, for example, for group1, we are 95 percent confident that the mean of time in hospital is between 4.75 and 4.9 days. Similar conclusions can also be drawn from the other findings. In addition, the difference in means

$\mu_1 - \mu_2$ for female and male patients across the three readmitted categories was computed. The results of these t.tests are shown in Table 4.

Table 2: Summary of point estimates - sample mean values

Female patients readmitted within (less than) 30 days	4.82 days
Female patients readmitted after 30 days	4. 59 days
Female patients who were not reported as readmitted	4.32 days
Male patients readmitted within (less than) 30 days	4.70 days
Male patients readmitted after 30 days	4.37 days
Male patients who were not reported as readmitted	4.17 days

Table 3: Summary of 95 percent interval bands for the six groups

Group 1 95 percent confidence interval: 4.750470 4.900051 sample estimates: 4.82526	Group 2 95 percent confidence interval: 4.553174 4.637112 sample estimates: 4.595143	Group 3 95 percent confidence interval: 4.293136 4.361248 sample estimates: 4.327192
Group 4 95 percent confidence interval: 4.617475 4.784254 sample estimates: 4.700865	Group 5 95 percent confidence interval: 4.328110 4.420377 sample estimates: 4.374243	Group 6 95 percent confidence interval: 4.136519 4.208910 sample estimates: 4.172714

Table 4: Summary of difference in means for three readmitted categories

Patients readmitted within 30 days 95 percent confidence interval: 0.01239246 0.23639859 sample estimates: 4.825260 4.700865	Patients readmitted after 30 days 95 percent confidence interval: 0.1585342 0.2832648 sample estimates: 4.595143 4.374243	Patients not reported as readmitted 95 percent confidence interval: 0.1047804 0.2041750 sample estimates: 4.327192 4.172714

E. Test of Significance

In the next part of the project, we carried out hypothesis testing. We perform hypothesis test to compare two population means μ_F, μ_M where *F: female patients , M: male patients*

$$H_o: \mu_F = \mu_M$$
$$H_1: \mu_F > \mu_M$$

Where H_1 implies that female patients on an average, spent more time in hospital compared to male patients. The analysis was undertaken through visualization, using boxplots, qq-plots and test of significance. The results for patients readmitted within 30 days of discharge are presented in this report. Similar analysis was conducted for other two categories as well. Figure 6 shows the box plot for patients readmitted within 30 days. Figure 7, on the other hand, represents the qq plots for female and male patients readmitted within 30 days. Since, we have a huge dataset with skewed distribution and the variable of significance, time in hospital is integer valued, the qq plots do not fall within the normal confidence bands. We conducted Welch two sample t-test as a test of significance for hypothesis testing. We used $\alpha = 0.05$ for the test. The results of the test for patients readmitted within 30 days of discharge are shown in Figure 8.

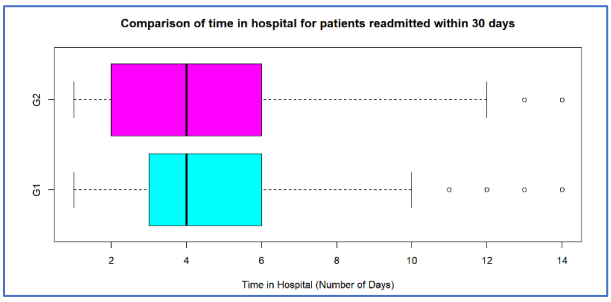

Figure 6: Graphical description of time in hospital for male and female patients using box plot visualization

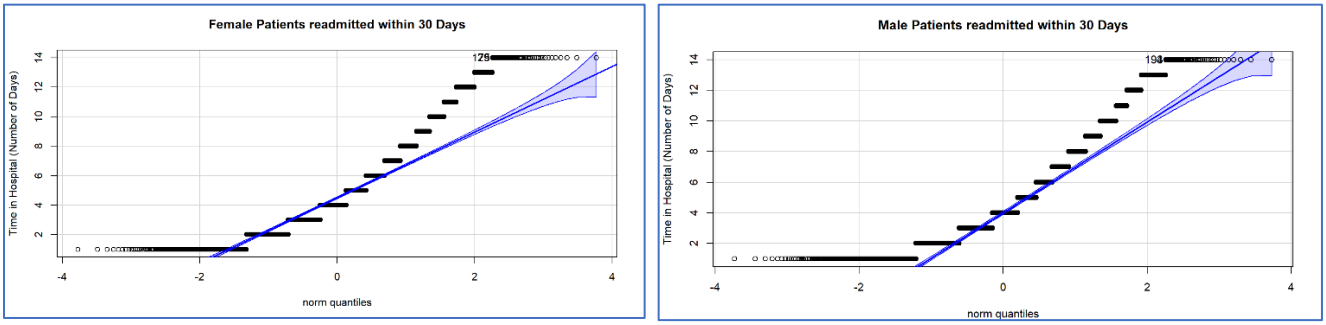

Figure 7: Visualization of data using qqplots for female patients (left) and male patients (right)

```
        Welch Two Sample t-test

data:  diabetic_data.lessF$time_in_hospital and diabetic_data.lessM$time_in_hospital
t = 2.1771, df = 10950, p-value = 0.01475
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.03040203        Inf
sample estimates:
mean of x mean of y
 4.825260  4.700865
```

Figure 8: Test results for Welch two sample t-test for patients readmitted within 30 days

The results indicate that the p-value of 0.01475 is small, showing that this evidence (difference between 4.82 and 4.70) would be unlikely if in fact $\mu_F = \mu_M$. Formally speaking, with p-value (0.01475) < α (0.05), we can reject H_o, which states that average length of stay for male patients and female patients who were readmitted within 30 days is the same. Standard deviation (s) for this group is 3.028 The t-value (2.1771) < 3s where s is the standard deviation for this sub-sample. As we reject H_o and H_1 is true, the results indicate correct decision. Statistically, average length of stay for female patients is observed to be greater than male patients. According to the dataset, length of stay in hospital is reported as number of days (integers). Practically speaking, the average length of stay of 4.32 and 4.17 days would have to be perceived in the form of hours (when the patient was discharged from the hospital) according to the context of the problem.

In essence, this report summarises the basic descriptive statistics, analysis of the length of stay of patients with respect to factors such as gender and readmission. In the consequent part of the project, we plan to study correlation and regression for different variables in order to predict readmissions and assess factors contributing to increased hospital readmissions.

References

Healthstream. (2021, April 1). Retrieved from The Economic & Emotional Cost of Hospital Readmissions: https://www.healthstream.com/resource/blog/the-economic-emotional-cost-of-hospital-readmissions

Hospital Readmissions. (2018). Retrieved from U.S. HHS, CMS, Office of Minority Health's Mapping Medicare Disparities Tool: https://www.americashealthrankings.org/learn/reports/2021-senior-report/appendix-measures-table

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*.

UCI Machine Learning Repository. (2014). *Center for Machine Learning and Intelligent Systems*. Retrieved from https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Wang, S., & Zhu, X. (2022). Predictive Modeling of Hospital Readmission: Challenges and Solutions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2975-2995.

Appendix A: Data Dictionary

S.No.	Field Name	Field Type	Description and Values
1	Encounter ID	Numeric	Unique identifier of an encounter
2	Patient number	Numeric	Unique Identifier of a patient
3	Race	Character	Values: Caucasian, Asian, African American, Hispanic, and other
4	Gender	Character	Values: male, female, and unknown/invalid
5	Age	Character	Grouped in 10-year intervals: [0, 10), [10, 20) ..., [90, 100)
6	Weight	Numeric	Weight in pounds
7	Admission Type	Character	Integer identifier corresponding to 8 distinct values, for example, emergency, urgent, elective, newborn, and not available
8	Discharge Disposition	Character	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, discharged/transferred to another type of inpatient care institution and not available
9	Admission source	Character	Integer identifier corresponding to 26 distinct values, for example, physician referral, clinic referral, court/law enforcement, emergency room, and transfer from a hospital
10	Time in Hospital	Numeric	Integer number of days between admission and discharge
11	Payer code	Character	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\BlueShield, Medicare, and self-pay
12	Medical Specialty	Character	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon
13	Number of lab procedures	Numeric	Number of lab tests performed during the encounter
14	Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter
15	Number of medications	Numeric	Number of distinct generic names administered during the encounter
16	Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter
17	Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter
18	Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
19	Diagnosis 1	Character	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
20	Diagnosis 2	Character	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
21	Diagnosis 3	Character	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
22	Number of Diagnoses	Numeric	Number of diagnoses entered to the system
23	Glucose serum test result	Character	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
24	Alc test result	Character	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal “if the result was less than 7%, and “none” if not measured.
25	Change of medications	Character	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
26	Diabetes medications	Character	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”

27	24 features for medications	Character	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was 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
28	Readmitted	Character	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.