QMST 5334 Final Report

**Analysis of hospital readmissions for patients with diabetes**

*Neeraj Kumar Reddy Panta, Ruchi Dilip Kukde*

1. **Introduction**

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 statistical analytical tools to identify underlying causes for readmissions, to achieve meaningful and transparent predictions for effective decision making. The report is divided into various sections describing purpose and description of dataset, basic statistics, descriptive plots, linear regression, analysis of residuals and analysis of variance.

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 grants 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 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:

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of the stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.
6. **Descriptive Analytics**

Exploratory analysis is the first and critical step in the data analysis process. It involves performing initial data investigations to discover patterns, spot anomalies, test hypothesis and check assumptions. It is done by describing 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’.

Graphical user interface, application, table, Excel

Description automatically generated

Figure 1: Illustration of ‘time\_in\_hospital’ variable from dataset(UCI Machine Learning Repository, 2014)

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 of time in hospital**

|  |  |
| --- | --- |
| **Measures of Location (in Days)** | |
| Mean | 4.395987 |
| Median | 4 |
| Mode | 3 |
| **Measures of spread (in Days)** | |
| Range | 13 |
| IQR | 4 |
| Variance | 8.910868 |
| Standard Deviation | 2.985108 |

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

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

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. 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. For further analysis we take a closer look at the number of medications administered and number of lab procedures carried out for the patients during their hospital stay. Figure 3 shows the distribution of number of medications administered to the patients in histogram and box plot representations. The histogram and box plot visual representations for number of lab procedures is shown in Figure 4. These figures indicate that the distributions are right skewed. Also, there are many outliers present in these distributions. From Table 2, it is noted that on an average 16 medicines were administer to the patients and 43 lab tests/ procedures were carried out. 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, how many medications were administered, were any lab procedures carried out?

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure 3: Graphical description of num of medications using histogram (left) and boxplot (right)

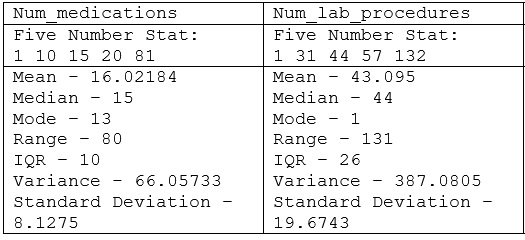
Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

Figure 4: Graphical description of num of lab procedures using histogram (left) and boxplot (right)

Table 2: Summary of basic statistics



1. **Analysis of Variance (ANOVA)**

Regression Analysis is a way of predicting the value of one variable from another. It is a hypothetical model of the relationship between two variables. The models we have used are linear ones. Therefore, we describe the relationship using the equation of a straight line. A model is used to predict an outcome (Y) based on set of predictors (X). For both the Simple and Multiple Regression Models we are considering the entire sample of the dataset of almost 100,000 samples.

Text

Description automatically generatedText

Description automatically generated In the simple regression model developed for this project, we wanted to see how the value of time\_in\_hospital (in days) varies for patients who have been re-admitted in hospital within 30 days based on the number of medications administered. Using outcome variable: time\_in\_hospital () And predictor variable: Num\_medications (), the equation of the model is: time\_in\_hospital = b1(num\_medications)+. Pearson’s and spearman correlation tests helped us to determine the relationship as shown in Figure 5. The correlation tests display estimate value with -1 to +1 so an estimate that is close to -1 is strong negative and the closer it gets to zero the weaker it gets and vice versa for the positive estimate. In our case we got both the test estimates to be around 0.48, which indicates a moderate positive correlation. By performing preliminary assessments using scatter plots we determine the correlation between the outcome and predictor variables. In the scatter plot, as depicted in Figure 6a, we cannot fully understand the relationship between the two variables time\_in\_hospital and num\_medications as the variables are measured on a discrete scale, this presents us with a challenge. However, there could still be a linear relationship difficult to deduce. So, by using the heat map R scatter plot along with some base R functions such as smooth scatter and contour functions based on defining colors to regions according to the density of data instances present, we obtain heatmap R scatter plot shown in Figure 6b. From the heat map it is a positive correlation as confirmed by the tests. We obtain the equation with intercept and slope coefficients as: *time\_in\_hospital = 0.17929 (num\_medications) + 1.73771*

Figure 5: Pearson’s and Spearman’s correlation test values

|  |  |
| --- | --- |
| *Chart, table  Description automatically generated*  Figure 6a: Scatter Plot to depict the correlation between outcome and predictor variables. | A computer screen capture  Description automatically generated with low confidence  Figure 6b: Heat map for depicting the correlation between the outcome and predictor variables. |

Text

Description automatically generated with medium confidenceText

Description automatically generatedFigure 7: Summarized Output of the Simple Linear Model

Figure 8: Confidence Interval for Slope

From Figure 7, we understand that the num\_medications coefficient suggests for every 10-count increase in the count of medications of the patient, we can expect increase in length of stay by 0.17929 \* 10 = 1.7929 days, on average. Is the Slope statistically significant? Yes. It is, because the value of slope () in the Interval doesn’t contain zero. Also, if the coefficient is large compared to the standard error, then statistically our coefficient will not be zero. From the 95% confidence interval test, shown in Figure 8, we can say with 95% confidence that slope lies between 0.1732 and 0.1853. Figure 5 also shows that num\_medications coefficient is 58.21 standard errors away from zero and it is far from zero. The larger our t-statistic, the more certain we can be that the coefficient is not zero. The p-value is calculated using t-statistic from the t distribution and helps us understand our coefficient's significance to the model. In practical terms any p-value below 0.05 is significant. In our model, we can see that the intercept and num\_medications have p-value of 2e-16 which is extremely small, and it is even below 0.001. We can conclude that the coefficients in this model are not zero. The residual standard error is a measure of how well the model fits the data. From our model summary, we can see that on average, the actual values are 2.658 days away from regression line on average. The F-statistic and overall p-value help us determine the result of our Hypothesis test. It is common for the F-statistic to be close to 1 if we have lot of predictors. However, for smaller models, a larger F-statistic and a small p-value indicates that null hypothesis should be rejected, and it clearly indicates that the coefficient in the model is not zero.

For Multiple Linear Regression model, we wanted to see how the value of time\_in\_hospital (in Days) varies for patients who have been re-admitted in hospital within 30 days based on the number of medications they have also used number of lab procedures they have undergone. In this case, The equation of the model: time\_in\_hospital = b1(num\_medications)+b2(num\_lb\_procedures)+ . We obtain the multicollinearity scatter plot shown in Figure 9. To quantify the correlation among the variables we use a function in R and obtain the correlation matrix as depicted in Figure 10. The equation with intercept and slope coefficients is obtained as 0.1578 (Num\_medications) + 0.0299 (Num\_lab\_procedures) + 0.7758. From the summarized view of multiple regression model shown in Figure 11, the p-value of F-statistic is <2.2e-16, which is highly significant, this means at least one of the predictor variables is significantly related to the outcome variable. Also, change in number of num\_medications and num\_lab\_procedures, the time\_in\_hospital of a patient is associated. For instance, for 10 count increase in the number of medications taken by the patient, we can expect an increase of 0.1578 \* 10 = 1.578 days of patient staying in hospital (when the num\_lab\_procedures are constant). The confidence interval for model coefficients is shown in Figure 12. What about the goodness of fit for the Model? The adjusted R-squared value for multiple regression model is 0.2627, meaning that 26.27% of the variance in measure of days can be predicted by num\_medications and num\_lab\_procedures number count. This model is better than the simple linear model with only num\_medications which had an adjusted R-squared of 0.2297. The RSE gives us a measure of error of prediction. The multiple linear regression model gives an RSE rate of 54%, which is better than the simple linear regression model where the RSE was 0.558 (i.e., 55.8% error rate).

Diagram

Description automatically generated

Figure 9: Multi Collinearity Scatter plot

Text

Description automatically generated

Figure 10: Correlation Matrix

Text

Description automatically generated

Figure 11: Summary of the Multiple Linear Regression Model

Graphical user interface, text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generatedFigure 12: Confidence Interval for the model coefficients

Figure 13: Residual Plot for the Simple Regression Model

Text

Description automatically generated For both models, we analyzed the residuals, to verify the Gauss Markov Theorem conditions. Additionally, we also try to explain and diagnose what multicollinearity is and explain how violations to the theorem affect the model. The residuals plot for the simple regression model is shown in Figure 13. From Simple linear regression Model, we used in regression analysis, we compute the mean value of residuals, i.e., -3.671336. Therefore, the first condition of Gauss Markov Theorem that errors have expectation zero is satisfied. Next, we test the residuals are not correlated among themselves, for this we do the Durbin Watson test. From Figure 14, the results of Durbin Watson test show that with a low p-value we reject the null hypothesis. We conclude that the residuals are correlated with each other. We then perform Breusch-Pagan test (Figure 15) to see if the Homoscedastic condition is satisfied or not satisfied. Based on the small p-value we reject the hypothesis that residuals are homoscedastic. We conclude that the errors do not have equal variances. We test for normality of the residuals by using Histogram (Figure 16) and QQ-plot (Figure 17). We observe that the distribution of errors is rightly skewed. Therefore, normality condition is not satisfied.​

Text

Description automatically generatedFigure 14: Results for Durbin Watson Test

Chart, histogram

Description automatically generatedFigure 15: Results for Breusch Pagan Test

Figure 16: Histogram to test normality

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedFigure 17: QQ-plot for the residuals

Chart, scatter chart

Description automatically generatedFigure 18: Residual plot for predictor variable Num\_medications

Figure 19: Residual Plot for predictor num\_lab\_procedures

From Figure 17, the qq-plot shows that few samples lie within the confidence band and by this we can state that the normality conditions are not satisfied. We try to test for normality Shapiro-Wilk test, if the residuals are Gaussian Distributed or not and since the number of samples is too large, we could not conduct this formal test. Therefore, from the above test we can say there is correlation between the variables we considered for the simple linear regression model. Similar analysis of residuals is carried out for multiple linear regression models. Figures 18 and 19 show the model's residuals with respect to the number of medications and lab procedures. For the multiple regression model, we computed the mean of residuals, i.e., 5.223827e-16, therefore the first condition of Gauss Markov theorem that errors have expectation of zero is satisfied.​ Next, we test to see if the residuals are correlated with each other or not, using Durbin Watson test, a low p-value is obtained, we reject null hypothesis. We conclude that the residuals are correlated with each other.​ Next, we perform Breusch-Pagan test to see if the homoscedasticity condition is satisfied or not.​ We reject the hypothesis that the residuals show homoscedasticity. Due to the p value being exceedingly small, we conclude that the errors do not have equal variances.

Chart, histogram

Description automatically generatedFigure 20: Histogram to check normality

Chart, line chart

Description automatically generated

Figure 21: QQ-plot of multiple linear regression model’s residuals to check normality

Using histogram shown in Figure 20, we conclude that there is no normality of the residuals as it is right skewed, and from the QQ-plot depicted in Figure 21, we conclude that the samples do not lie in the confidence band which results in not being normal. As seen from the simple linear regression model, the Shapiro-Wilk test does not work for the multiple linear regression model as well since the number of samples we have been working on for the model is large. Consequently, we checked for the multicollinearity between the variables, which indicates whether the predictors are correlated. If the predictors are correlated, the estimation of regression coefficients gets affected which indicates multicollinearity!​

Diagram

Description automatically generated

Figure 22: Scatter plots for three variables to test multicollinearity

Text, letter

Description automatically generatedFigure 23: Correlation Matrix for the Multiple linear regression model

From Figure 22, we can see the trend of the collinearity among the three variables, however using the correlation function, we try to quantify the correlation which is said to be positive if the estimate is greater than 0. As displayed in Figure 23, we observe that the predictors number of medications and number of lab procedures are correlated. The square root of the VIF (Variance Inflation Factor) says how much larger the standard error is, as compared to the case of predictors that are uncorrelated. So, we calculated the VIF for the model which is obtained as 1.099588. Our model has VIF > 1 which indicates there is no reason for concern, and it is possible to say that there is correlation between the two predictors. Looking at the results of both models, it is safe to say that regression models 1 and 2 satisfy only one of the four conditions of the Gauss-Markov theorem conditions, i.e., expectation of errors is zero​. From the tests conducted it is visible that residuals are correlated and display heteroscedasticity and are not normally distributed. ​ Since the number of samples is greater than 5000, the formal Shapiro-Wilk test cannot be performed on our data.​ There is evidence of multicollinearity among the predictors used in regression model 2.​

Surface chart

Description automatically generated with medium confidence With the results obtained from regression testing and analysis of residuals it is evident that analyzing models with continuous variables or numerical data is not always an appropriate modelling approach. With the help of the methodologies used in analysis of variance, we can use categorical variables. We create models by factoring the categorical variables and adding them to the linear models.

*Text, letter

Description automatically generated*Figure 24: Heat scatter plot for displaying correlation between the variables

Figure 25: Summarized Output of the Simple Linear Model

A picture containing text

Description automatically generated By performing analysis of variance (ANOVA) we can determine up to what percent of the samples have been explored, and how much are yet to be extracted and explained, and instead of creating multiple t-tests we use ANOVA groups differences by comparing means of each group. Like a t-test, we use ANOVA test with a null hypothesis that the means are the same. Since we can know the method that will let us use categorical variable in our linear models, we are designing new models with random 500 samples from the total 100,000 samples. We are considering the same simple linear regression model as in the before section, i.e., outcome variable is time\_in\_hospital () And predictor variable is num\_medications (). The equation of the model: = + .From Figure 24, we show a density heat scatter plot to understand correlation using R programming base functions such as Smooth scatter and contour. We computed Pearson’s coefficient which resulted in sample estimate of 0.40187 and 95% confidence intervals as 0.32 and 0.47. These results show that there is a positive correlation.

A screenshot of a computer

Description automatically generated with low confidenceFigure 26: 95% confidence intervals

Figure 27: Summarized Output of the Multiple Regression Linear Model

Chart, scatter chart

Description automatically generated From the summary shown in Figure 25, we can understand that the slope is significant since it is not equal to zero, and t value is not small. Additionally, the p-value is highly significant when we consider the alpha to be 95% which is 0.05. We check the 95% confidence interval for the values of intercept and slope, as shown in Figure 26. The goodness of fit of the model is 0.6204048, and with all the values and interpretation we can assume the model to be a good one. For the multiple linear regression model the outcome variable is same as the previous model i.e., time\_in\_hospital, but this time in the predictor variables instead of just having numerical variable we have taken an addition of a categorical variable. The outcome variable time\_in\_hospital (). The predictors variables are num\_medications () & readmitted (X2). The equation of the model: = + + . From the summary shown in Figure 27, for the new multiple regression model, we can see that both the predictor variables are significant, since the estimate coefficients are not equal to zero, but when analyzing the t-value the readmittedNO is closer to zero since it is 0.424, and the rest three are 2 and above. The p values are significant, but it varies based on the alpha we assume for the study. Since the alpha we take is 0.05, only readmitted<30 and readmitted>30 is significant and not readmittedNO.

Figure 28: Scatter plot based on groups of the categorical variable

Chart, scatter chart

Description automatically generated

Figure 29: Linear regression analysis predicted lines

Timeline

Description automatically generated

Figure 30: Histogram to visualize time in hospital based on categorical variables

When comparing the multiple models with linear model, the adjusted R-squared is better as 0.1676 > 0.1598. Also, the residual standard error is less for the multiple regression model when compared to the simple model. So, we can say that the new multiple linear regression model is better than the simple model. Figure 28 shows the scatter plot of the three groups of the categorical variable with respect to number of medications and number of days in the hospital. It is interesting to note that most of the samples lie between number of days around 1-10 days and number of medications administered as 1-30. Using linear regression analysis, the predicted lines are obtained and shown in Figure 29. To check the variability in each group, we plot the histograms of the categories separately. The histograms are shown in Figure 30. The box plots are shown in Figure 31 to explain the variability of observations in each group.

We designate readmitted<30 as the base group and outcome variable as time\_in\_hospital. The equation of the regression model considering the categorical variables is: . The results obtained for ANOVA are depicted in Figure 32. It is observed that intercept (base category) is significant as compared to readmittedNO and readmitted>30. In addition, this model shows that a low R squared value id obtained for this categorical variable as a predictor for time in hospital. The amount of variability is explained by F-ratio. In this case, the F-ratio is 3.43. It is an indicator of amount of variability explained by the model and compared to the error in the model. Since the F-ratio is high, it means that there is a difference in means. Similar conclusions can be drawn from the summary of the ANOVA analysis from Figure 33. Since the ANOVA reflects whether the experiment was successful or not, additional tests are required to find out where the differences lie. We use Tukey post-hoc test to compare all pairs of means as shown in Figure 34. In the Tukey Post-hoc tests, a single-step multiple comparison procedure is followed to find means that are significantly different from each other. It compares all pairs of means. It is based on a studentized range distribution (q). From the Tukey post-hoc tests and confidence intervals, there is a difference in the means of the different groups. Figure 35 shows the confidence intervals for multiple comparisons of means using Tukey contrasts. In conclusion, we can perform analysis of variance on the outcome variable time in hospital. It would be interesting to test our model for logistic regression as a part of further investigation.

Chart, box and whisker chart

Description automatically generatedFigure 31: Visual comparison using boxplots to account for variability of observations within each group

Text

Description automatically generated

Figure 32: Summary of regression model (ANOVA)

A picture containing calendar

Description automatically generated

Figure 33: Summary of regression model (ANOVA)

A screenshot of a computer

Description automatically generated with medium confidence

Figure 34: Tukey Post Hoc tests for the model

A picture containing text

Description automatically generated

Figure 35: Tukey Post Hoc confidence intervals

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 post mid-term part of the project, we studied the correlations and performed regression analysis for different variables. We conducted the analysis of residuals and analysis of variance for the dataset. In conclusion, this project was a great learning opportunity for statistical methodology for real-life application.

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, Nate glinide, 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. |