Predicting readmission probability for diabetes inpatients

STAT 471/571/701, Fall 2017

Cheshta Dhingra

Due: April 7, 2017 at 11:59PM

I. Executive Summary

a. Background

Diabetes is a chronic medical condition affecting millions of Americans, but if managed well, with good diet, exercise and medication, patients can lead relatively normal lives. However, if improperly managed, diabetes can lead to patients being continuously admitted and readmitted to hospitals. Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. As a result, the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge.

Given these policy changes, being able to identify and predict those patients most at risk for costly readmissionshas become a pressing priority for hospital administrators. In this project, we shall explore how to use the techniques we have learned in order to help better manage diabetes patients who have been admitted to a hospital. Our goal is to avoid patients being readmitted within 30 days of discharge, which reduces costs for the hospital and improves outcomes for patients. The goals of this analysis are: 1. Identify the factors predicting whether or not the patient will be readmitted within 30 days. 2. Propose a classification rule to predict if a patient will be readmitted within 30 days.

b. Data

The original data is from the Center for Clinical and Translational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this dataset, from ~70,000 unique patients. The data includes demographic elements, such as age, gender, and race, as well as clinical attributes such as tests conducted,emergency/inpatient visits, etc. All observations have five things in common:

1. They are all hospital admissions
2. Each patient had some form of diabetes
3. The patient stayed for between 1 and 14 days.
4. The patient had laboratory tests performed on him/her.
5. The patient was given some form of medication during the visit.

A more detailed summary of each variable can be found below in section IIA.

c. Methods

To developed my final model I created models using two different methods: backward selection and LASSO. I attempted to use AIC as a third method but found that it was too computationally expensive for my computers and so I abandoned this method in favor of the other two. I then compared the ROC curves and AUC for the three different models and chose the one I felt had the most predictive power based on the ROC curves, AUC and significance of included variables. I then found the associated misclassification error using the given fact that the cost of mislabelling a readmission is twice the cost of mislabelling a non-readmission. This gave me my final Bayes Rule Classification Threshold. Finally, I tested my model on my subset of test data.

d. Main Findings

The key variables of importance were found to be: num_procedures, num_medications, number_emergency, number_inpatient, A1Cresult, metformin, glimepiride, insulin, diabetesMed, disch_disp_modified, adm_src_mod, age_mod, diag1_mod, diag2_mod, diag3_mod. The Bayes Rule Classification Threshold using the risk ratio of 2:1 was 1/3. This means that if the predicted probability of readmission exceeds 1/3, we will predict that that individual gets readmitted. Our misclassification error was about 22%.

e. Limitations

Readmissions is not a flawless indicator of hospital quality. Some of the advantages of this measure are: it has been shown that a high readmission rate suggests poor care, it has good face validity, is relatively easy to identify and we have a large amount of data regarding readmissions. However, it can be difficult to predict, is often shown to lack association with other hospital characteristics associated with quality of care. Most importantly, whether or not a patient is readmitted may be associated with factors outside the hospital's control, such as the patient's medication compliance level after she is discharged.
Further, it may be interesting to see the 60 day and 90 day readmission data.

II. Data Analysis

a. Data Summary

Description of variables

The full dataset used covers ~50 different variables to describe every hospital diabetes admission. In this section we give an overview and brief description of the variables in this dataset.

a) Patient identifiers:

a. `encounter_id`: unique identifier for each admission
b. `patient_nbr`: unique identifier for each patient

b) Patient Demographics:

`race`, `age`, `gender`, `weight` cover the basic demographic information associated with each patient. `Payer_code` is an additional variable that identifies which health insurance (Medicare /Medicaid / Commercial) the patient holds.

c) Admission and discharge details:

a. `admission_source_id` and `admission_type_id` identify who referred the patient to the hospital (e.g. physician vs. emergency dept.) and what type of admission this was (Emergency vs. Elective vs. Urgent).
b. `discharge_disposition_id` indicates where the patient was discharged to after treatment.

d) Patient Medical History:

a. `num_outpatient`: number of outpatient visits by the patient in the year prior to the current encounter
b. `num_inpatient`: number of inpatient visits by the patient in the year prior to the current encounter

c. `num_emergency`: number of emergency visits by the patient in the year prior to the current encounter

e) Patient admission details:

a. `medical_specialty`: the specialty of the physician admitting the patient
b. `diag_1`, `diag_2`, `diag_3`: ICD9 codes for the primary, secondary and tertiary diagnoses of the patient. ICD9 are the universal codes that all physicians use to record diagnoses. There are various easy to use tools to lookup what individual codes mean (Wikipedia is pretty decent on its own)
c. `time_in_hospital`: the patient's length of stay in the hospital (in days)
d. `number_diagnoses`: Total no. of diagnosis entered for the patient
e. `num_lab_procedures`: No. of lab procedures performed in the current encounter
f. `num_procedures`: No. of non-lab procedures performed in the current encounter
g. `num_medications`: No. of distinct medications prescribed in the current encounter

f) Clinical Results:

a. `max_glu_serum`: indicates results of the glucose serum test
b. `A1Cresult`: indicates results of the A1c test

g) Medication Details:

a. `diabetesMed`: indicates if any diabetes medication was prescribed
b. `change`: indicates if there was a change in diabetes medication
c. `24 medication variables`: indicate whether the dosage of the medicines was changed in any manner during the encounter

h) Readmission indicator:

Indicates whether a patient was readmitted after a particular admission. There are 3 levels for this variable: "NO" = no readmission, "< 30" = readmission within 30 days and "> 30" = readmission after more than 30 days. The 30 day distinction is of practical importance to hospitals because federal regulations penalize hospitals for an excessive proportion of such readmissions. I regrouped this variable into an indicator for whether or not the patient was readmitted within 30 days.

See appendix for a summary of the variables in the full dataset.

Data cleaning

Since many of the values are missing, we will modify the dataset in the following ways:

1) `Payer code`, `weight` and `Medical Specialty` are not included since they have a large number of missing values.

2) Variables such as **`acetohexamide`**, **`glimepiride.pioglitazone`**, **`metformin.rosiglitazone`**, **`metformin.pioglitazone`** have little variability, and are as such excluded. This also includes the following variables: **`chlorpropamide`**, **`acetohexamide`**, **`tolbutamide`**, **`acarbose`**, **`miglitor`**, **`troglitazone`**, **`tolazamide`**, **`examide`**, **`citoglipton`**, **`glyburide.metformin`**, **`glipizide.metformin`**, and **`glimepiride.pioglitazone`**.

3) Some categorical variables have been regrouped. For example, `Diag1_mod` keeps some original levels with large number of patients and aggregates other patients as `others`.

4) Observations for which `race` was given to be "?" were omitted for ease of analysis. Similar process for observation where gender was "Unknown/Invalid". In total, 2276 observations were omitted, representing 2.2% of the dataset.

5) To ease in the analysis we removed `encounter_id` and `patient_nbr` as the large number of various values created issues and provided little predictive power. We are left with 29 variables in the dataset.

6) The event of interest is **readmitted within $<$ 30 days**. We have recoded those who were readmitted *beyond* 30 days such that they do not get counted under our event of interest.

Graphical Summary (See Appendix)

b. Analyses

Creating a Test set

I retained 10% of the data (about 10,000 observations) as a testing set to assess my final model. This leaves almost 90,000 observations to train the model.

Backward Selection

The first method I will use is Backward Selection. Starting with the full model I successively remove the variable with the highest p-value (lowest significance), then run the logistic regression with the remaining variables. This process is repeated until I reach a model which has only variables significant at the 0.05 level. The best model is shown below with its associated Anova test results (`fit_b.best`).

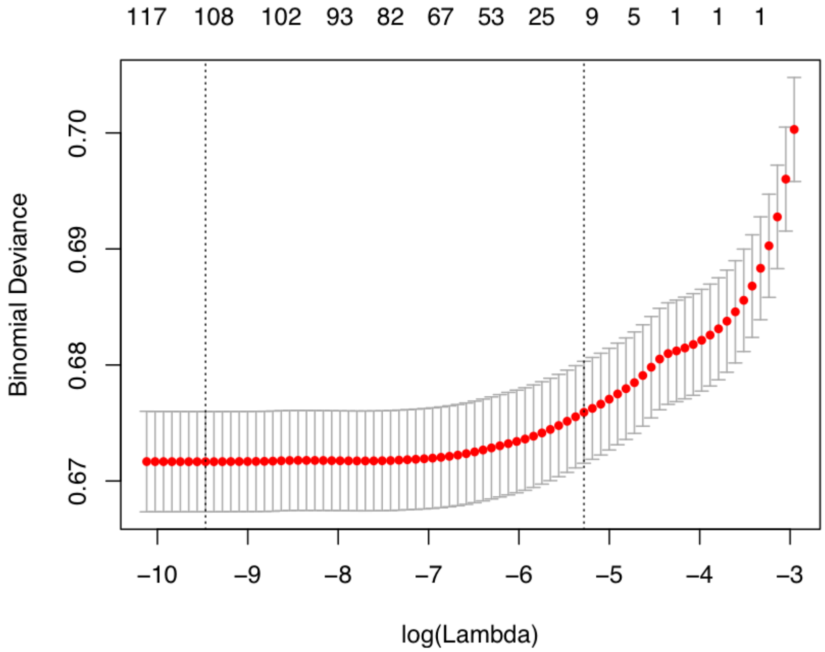
```
fit_b.best <- glm(read ~ num_procedures + num_medications + number_emergency +
    number_inpatient + A1Cresult + metformin + glimepiride + insulin + diabetesMed +
    disch_disp_modified + adm_src_mod + age_mod + diag1_mod + diag2_mod + diag3_mod,
    rtrain_clean, family = binomial)
Anova(fit_b.best)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: read
##                     LR Chisq Df Pr(>Chisq)
## num_procedures          5.86  1  0.0154571 *
## num_medications        19.04  1  1.278e-05 ***
## number_emergency       14.68  1  0.0001272 ***
## number_inpatient     1217.53  1  < 2.2e-16 ***
## A1Cresult              10.56  3  0.0143452 *
## metformin              17.50  3  0.0005569 ***
## glimepiride             8.32  3  0.0398108 *
## insulin                10.74  3  0.0132313 *
## diabetesMed            35.77  1  2.226e-09 ***
## disch_disp_modified   228.17  3  < 2.2e-16 ***
## adm_src_mod            14.25  3  0.0025888 **
## age_mod                38.11  3  2.674e-08 ***
## diag1_mod             179.04 23  < 2.2e-16 ***
## diag2_mod              62.59 24  2.731e-05 ***
## diag3_mod              89.87 20  7.798e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

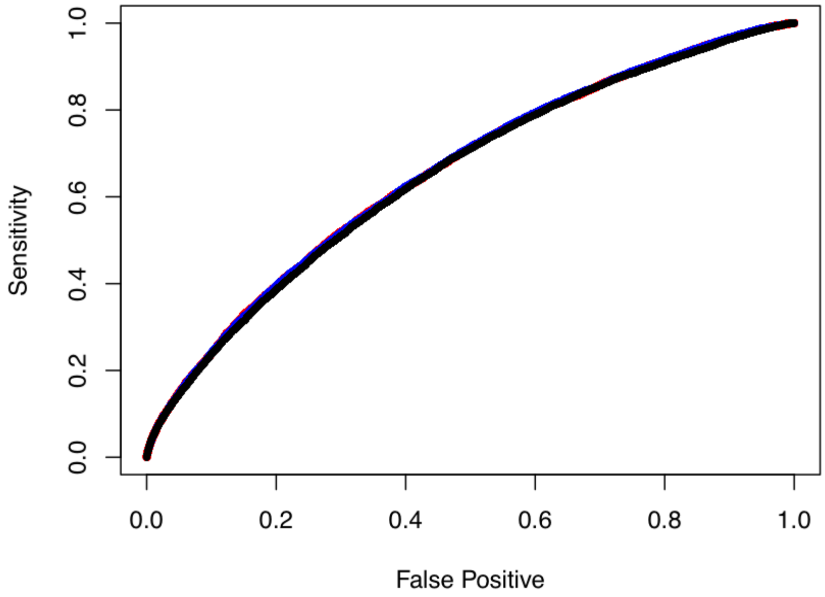
LASSO in classifications:

Next, I will use LASSO to find the best model variables. The regularization techniques used in regression are readily applied to classification problems. For a given lambda we minimize -log liklihood/n + lambda |beta| To remain consistent in both binary and continuous responses, glmnet() uses the following penalized least squares. RSS/(2n) + lambda |beta|

Shown below is the plot of the binomial deviance from the 10-fold cross validated LASSO model. We want to minimize this.



ackwards Selection, Blue: LASSO Lambda Min, Black: LASSO La



```
## Area under the curve: 0.6551
```

```
## Area under the curve: 0.6561
```

```
## Area under the curve: 0.6514
```

auc(fit_backward.roc) = 0.6551 auc(fit_lasso_min.roc) = 0.6561
auc(fit_lasso_1se.roc) = 0.6514

As we can see, all three models are fairly similar, with similar AUCs but I decided to go with the backward selection model since all of the variables in it are significant at the 0.05 level, which cannot be said about the LASSO models.

Now that we have selected our final model of read ~ num_procedures + num_medications + number_emergency + number_inpatient + A1Cresult + metformin + glimepiride + insulin + diabetesMed + disch_disp_modified + adm_src_mod + age_mod + diag1_mod + diag2_mod + diag3_mod

which was obtained through backward selection, we now need to come up with a reasonable classifier for our model. Based on a quick and somewhat arbitrary guess, it's estimated that it costs twice as much to mislabel a readmission than it does to mislabel a non-readmission. Based on this risk ratio, I will propose a specific classification rule to minimize the cost. Then:

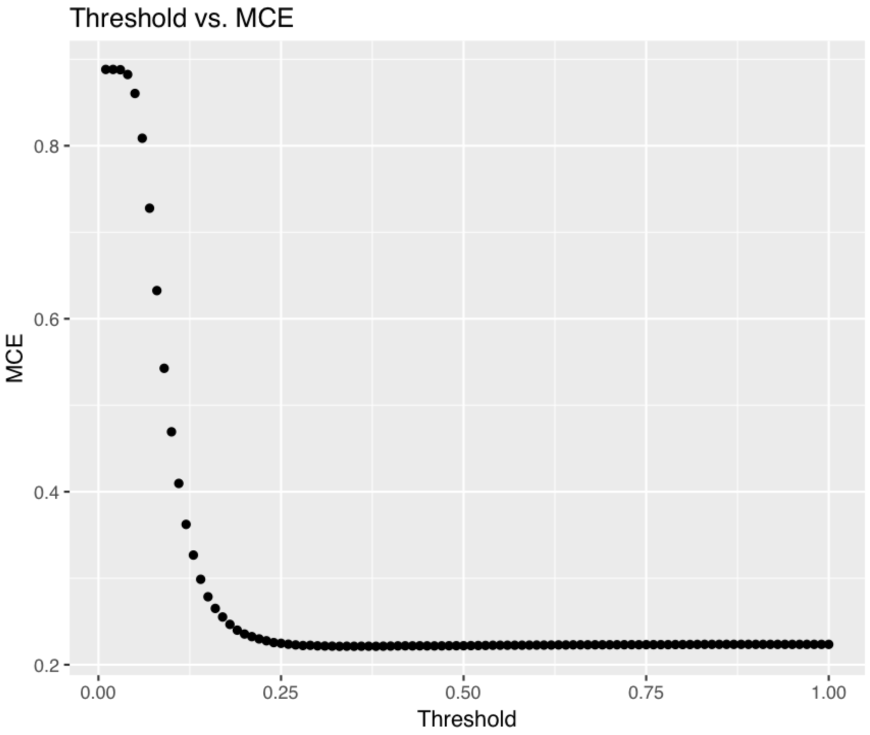
$a_{10}/a_{01} = 2 -> a_{01}/a_{10} = 1/2$

$P(Y = 1|x) > \frac{a_{01}/a_{10}}{1+a_{01}/a_{10}} = \frac{1/2}{1+(1/2)} = 1/3$

$logit > log(\frac{1/3}{1-1/3}) = log(1/2) = -0.693$ gives us the Bayes rule!

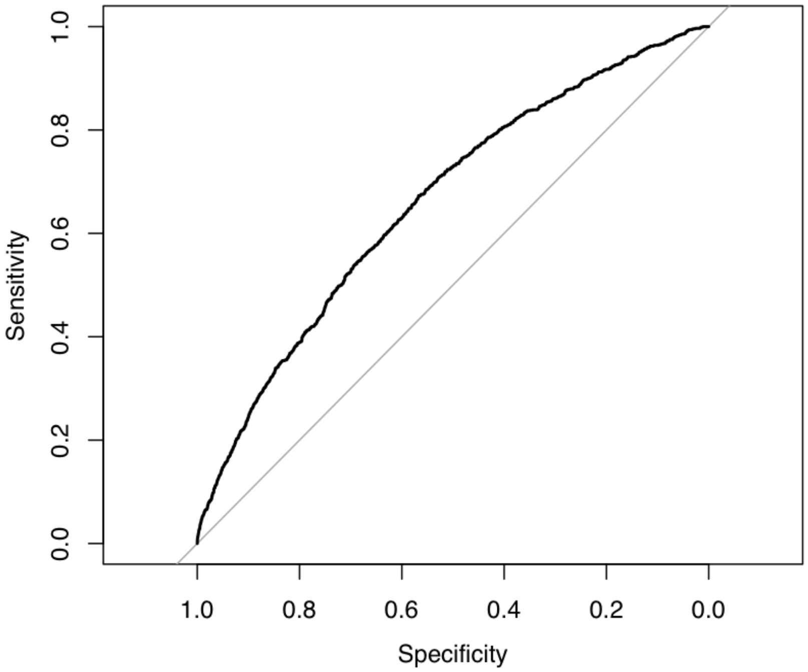
Therefore the Bayes Rule Classification Threshold using this risk ratio would be 1/3. Using this threshold we get a weighted misclassification error (MCE) of 0.22.

```r
ggplot(MCE.bayes.2, aes(x = Threshold, y = MCE)) + geom_point() + labs(title = "Threshold vs. MCE",
    x = "Threshold", y = "MCE")
```



Evaluating my model using testing data

Get the fitted prob's using the testing data:



```
## Area under the curve: 0.6605
```

c. Conclusion

The key variables of importance were found to be: num_procedures, num_medications, number_emergency, number_inpatient, A1Cresult, metformin, glimepiride, insulin, diabetesMed, disch_disp_modified, adm_src_mod, age_mod, diag1_mod, diag2_mod, diag3_mod. The Bayes Rule Classification Threshold using the risk ratio of 2:1 was 1/3. This means that if the predicted probability of readmission exceeds 1/3, we will predict that that individual gets readmitted. Our misclassification error was about 22%.

III. Citation

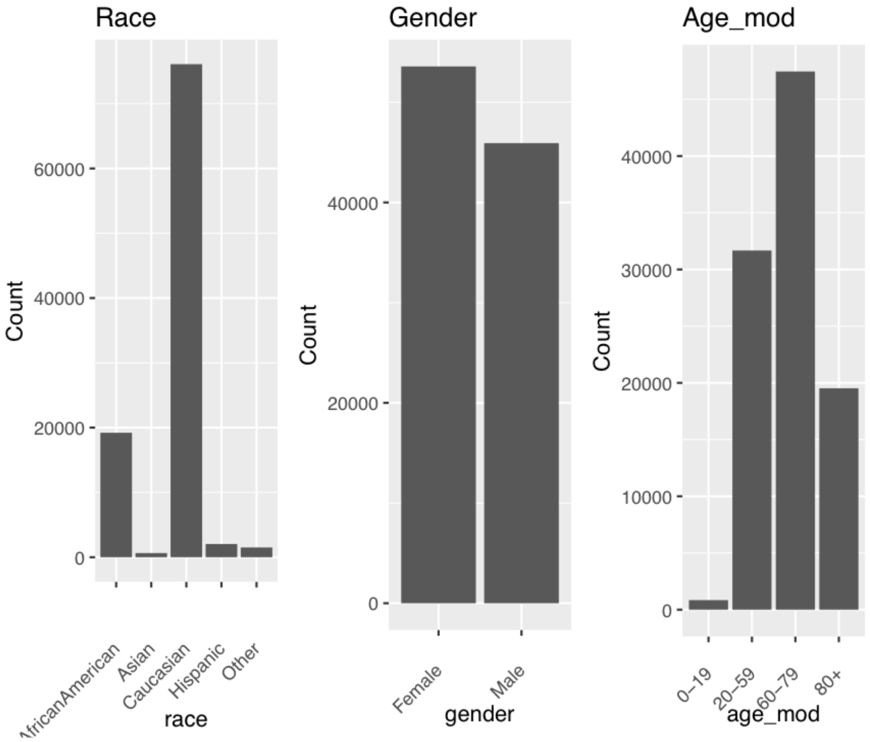
Data obtained from: [Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.] (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008) # IV. Appendix

Full dataset summary:

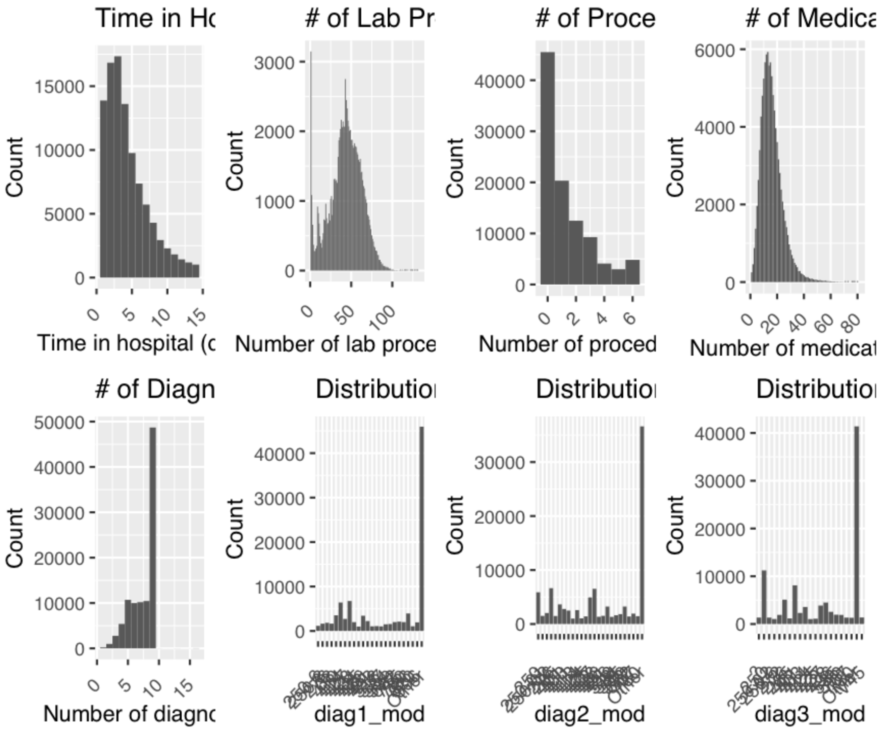
```
##   encounter_id         patient_nbr                     race      
##  Min.   :    12522   Min.   :       135   ?              : 2273  
##  1st Qu.: 84961194   1st Qu.: 23413221   AfricanAmerican:19210  
##  Median :152388987   Median : 45505143   Asian          :  641  
##  Mean   :165201646   Mean   : 54330401   Caucasian      :76099  
##  3rd Qu.:230270888   3rd Qu.: 87545950   Hispanic       : 2037
##  Max.   :443867222   Max.   :189502619   Other          : 1506
##
##              gender           age             weight
##  Female         :54708   [70-80):26068   ?        :98569
##  Male           :47055   [60-70):22483   [75-100) : 1336
##  Unknown/Invalid:    3   [50-60):17256   [50-75)  :  897
##                          [80-90):17197   [100-125):  625
##                          [40-50): 9685   [125-150):  145
##                          [30-40): 3775   [25-50)  :   97
##                          (Other): 5302   (Other)  :   97
##  admission_type_id discharge_disposition_id admission_source_id
##  Min.   :1.000     Min.   : 1.000           Min.   : 1.000
##  1st Qu.:1.000     1st Qu.: 1.000           1st Qu.: 1.000
##  Median :1.000     Median : 1.000           Median : 7.000
##  Mean   :2.024     Mean   : 3.716           Mean   : 5.754
##  3rd Qu.:3.000     3rd Qu.: 4.000           3rd Qu.: 7.000
##  Max.   :8.000     Max.   :28.000           Max.   :25.000
##
##  time_in_hospital   payer_code                 medical_specialty
##  Min.   : 1.000   ?      :40256   ?                     :49949
##  1st Qu.: 2.000   MC     :32439   InternalMedicine      :14635
##  Median : 4.000   HM     : 6274   Emergency/Trauma      : 7565
##  Mean   : 4.396   SP     : 5007   Family/GeneralPractice: 7440
##  3rd Qu.: 6.000   BC     : 4655   Cardiology            : 5352
##  Max.   :14.000   MD     : 3532   Surgery-General       : 3099
##                   (Other): 9603   (Other)               :13726
##  num_lab_procedures num_procedures num_medications number_outpatient
##  Min.   :  1.0      Min.   :0.00   Min.   : 1.00   Min.   : 0.0000
##  1st Qu.: 31.0      1st Qu.:0.00   1st Qu.:10.00   1st Qu.: 0.0000
##  Median : 44.0      Median :1.00   Median :15.00   Median : 0.0000
##  Mean   : 43.1      Mean   :1.34   Mean   :16.02   Mean   : 0.3694
##  3rd Qu.: 57.0      3rd Qu.:2.00   3rd Qu.:20.00   3rd Qu.: 0.0000
##  Max.   :132.0      Max.   :6.00   Max.   :81.00   Max.   :42.0000
##
##  number_emergency  number_inpatient      diag_1          diag_2
##  Min.   : 0.0000   Min.   : 0.0000   428    : 6862   276    : 6752
##  1st Qu.: 0.0000   1st Qu.: 0.0000   414    : 6581   428    : 6662
##  Median : 0.0000   Median : 0.0000   786    : 4016   250    : 6071
##  Mean   : 0.1978   Mean   : 0.6356   410    : 3614   427    : 5036
##  3rd Qu.: 0.0000   3rd Qu.: 1.0000   486    : 3508   401    : 3736
##  Max.   :76.0000   Max.   :21.0000   427    : 2766   496    : 3305
##                                      (Other):74419   (Other):70204
##      diag_3      number_diagnoses max_glu_serum A1Cresult
##  250    :11555   Min.   : 1.000   >200: 1485    >7  : 3812
##  401    : 8289   1st Qu.: 6.000   >300: 1264    >8  : 8216
##  276    : 5175   Median : 8.000   None:96420    None:84748
##  428    : 4577   Mean   : 7.423   Norm: 2597    Norm: 4990
##  427    : 3955   3rd Qu.: 9.000
##  414    : 3664   Max.   :16.000
##  (Other):64551
##   metformin     repaglinide     nateglinide     chlorpropamide
##  Down  :  575   Down  :    45   Down  :    11   Down  :     1
##  No    :81778   No    :100227   No    :101063   No    :101680
##  Steady:18346   Steady:  1384   Steady:   668   Steady:    79
##  Up    : 1067   Up    :   110   Up    :    24   Up    :     6
##
##
##
##  glimepiride    acetohexamide    glipizide      glyburide
##  Down  :  194   No     :101765   Down  :  560   Down  :  564
##  No    :96575   Steady:      1   No    :89080   No    :91116
##  Steady: 4670                    Steady:11356   Steady: 9274
##  Up    :  327                    Up    :  770   Up    :  812
##
##
##
##  tolbutamide     pioglitazone   rosiglitazone    acarbose
##  No    :101743   Down  :  118   Down  :   87   Down  :     3
##  Steady:    23   No    :94438   No    :95401   No    :101458
##                  Steady: 6976   Steady: 6100   Steady:   295
##                  Up    :  234   Up    :  178   Up    :    10
##
##
##
##    miglitol      troglitazone     tolazamide     examide     citoglipton
##  Down  :     5   No    :101763   No    :101727   No:101766   No:101766
##  No    :101728   Steady:     3   Steady:    38
##  Steady:    31                   Up    :     1
##  Up    :     2
##
##
##
##    insulin      glyburide.metformin glipizide.metformin
##  Down  :12218   Down  :     6       No    :101753
##  No    :47383   No    :101060       Steady:    13
##  Steady:30849   Steady:   692
##  Up    :11316   Up    :     8
##
##
##
##  glimepiride.pioglitazone metformin.rosiglitazone metformin.pioglitazone
##  No    :101765           No    :101764           No    :101765
##  Steady:     1           Steady:     2           Steady:     1
##
##
##
##
##
##  change     diabetesMed readmitted
##  Ch:47011   No :23403   <30:11357
##  No:54755   Yes:78363   >30:35545
##                         NO :54864
##
##
##
##
```

Here is a graphical summary of the 28 input variables and the response variable we retained in the working dataset `rdata_clean`.

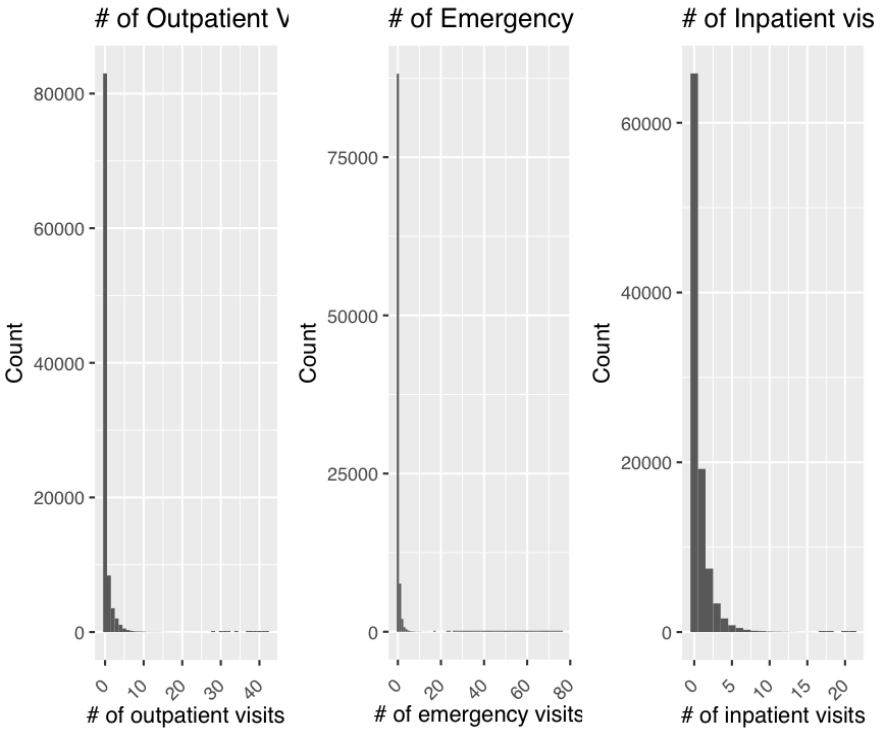
Distributions of Demographics



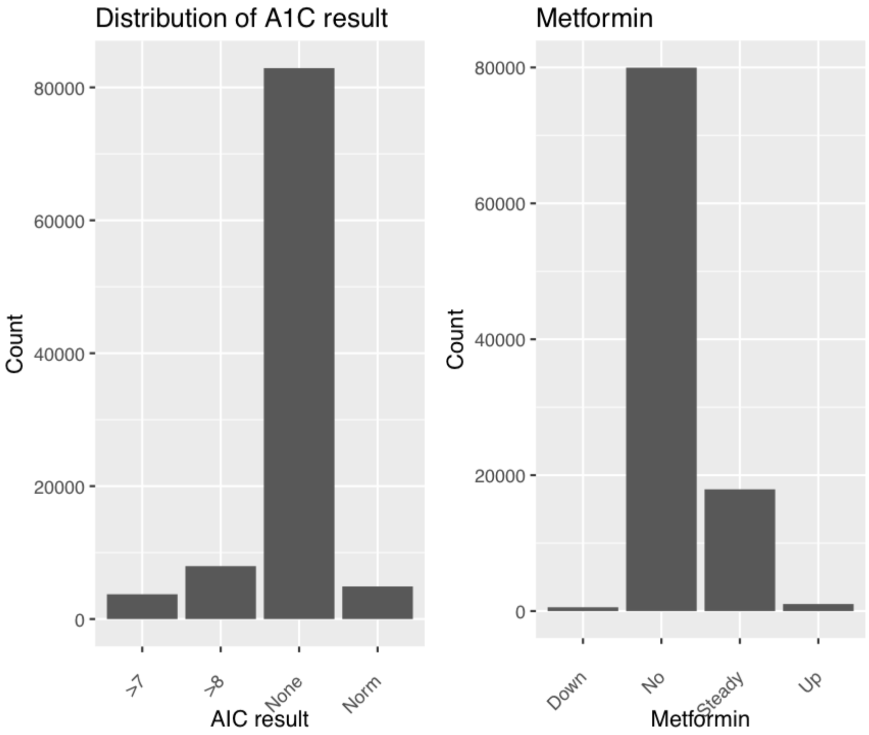
Distributions of Admission Details



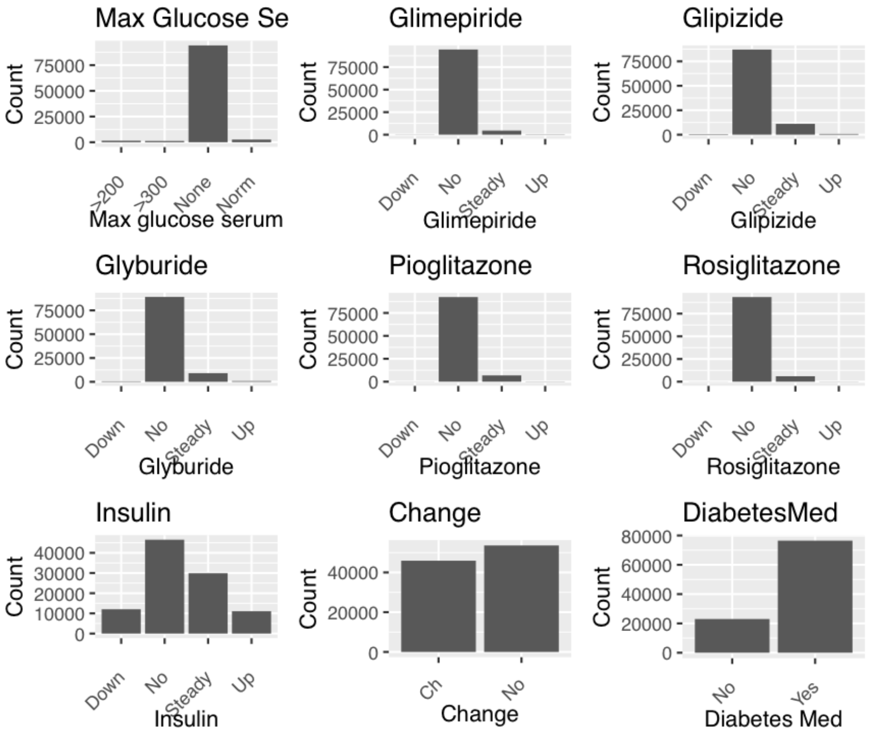
Distributions of Medical History variables



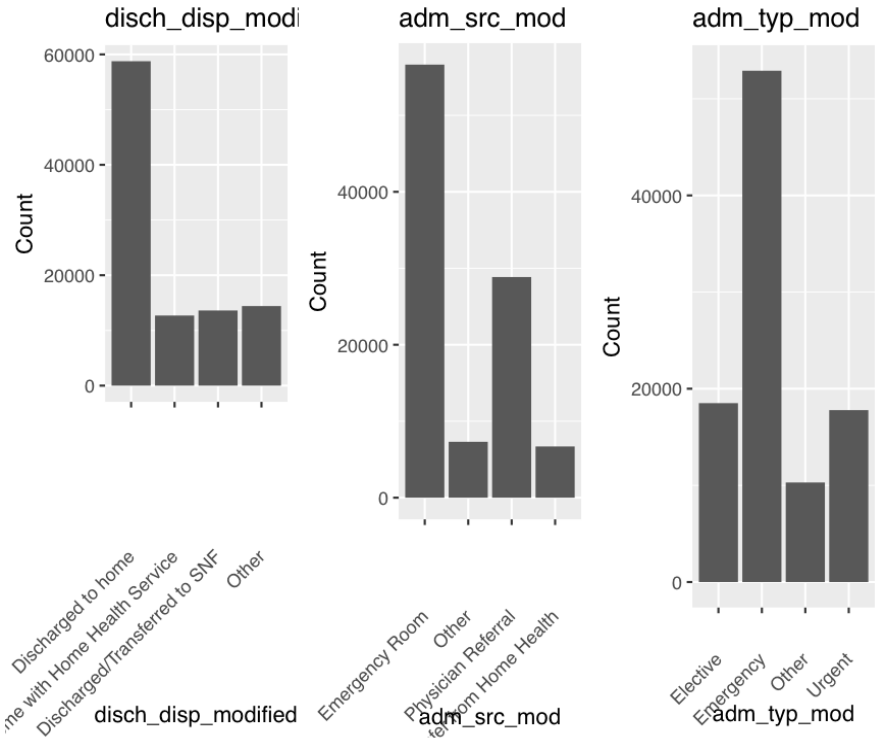
Distribution of Clinical Results variables



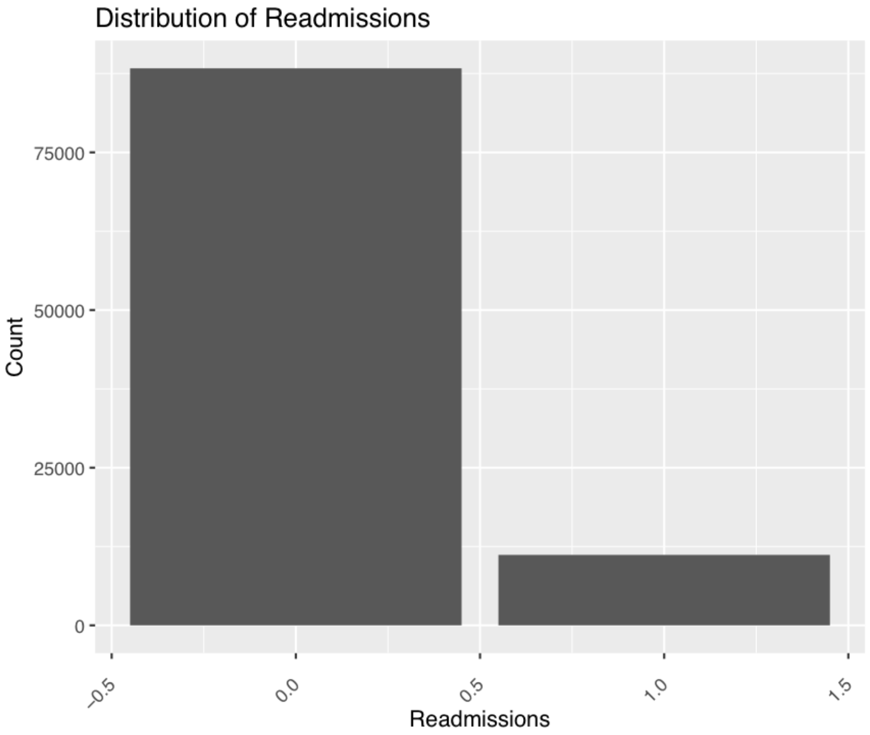
Distribution of Medication details variables



Distribution of admission/discharge details variables



Distribution of response variable (readmissions)


Distribution of Readmissions
Count
75000
50000
25000
0
–0.5
0.0
0.5
1.0
1.5
Readmissions