EMR Predictive Models For Patients with Diabetes

Keith Engwall

4/30/2018

Introduction

Electronic Medical Records (EMRs) can serve as a rich source of patient data, with which to construct predictive models for disease. Type 2 Diabetes accounts for 90-95% of all diabetes cases, with 23 million diagnosed cases among adults and 7.2 million undiagnosed cases in the United States as of 2015. ref1 Predictive models can provide a means to identify undiagnosed or at risk individuals for Type 2 Diabetes. This project develops and evaluates predictive models for identifying patients with diabetes using an EMR dataset.

Data

The project makes use of a 2012 EMR data set of approx. 10,000 patients, taken from a Kaggle competition ref2. The data set consists of sql tables for patients, allergies, diagnoses, and prescriptions, as well as tables for transcripts of visit data and labs. Figure 1 shows the relationship diagram between the Patient and Diagnoisis tables. Although all data relating to Type 2 Diabetes has been scrubbed from the dataset, patients diagnosed with Type 2 Diabetes are identified using an indicator in the patient table (not shown in the ER diagram). The full ER Diagram is available in Appendix A. Tables of interest are described below.

Patient Table

The patient table contains an indicator field to identify the diabetic population, as well as the patient's gender and year of birth

Allergy Table

The allergy table contains a field for allergy type to identify the category of allergy, as well as for reaction name for the type of reaction and severity name for the severity of the allergic reaction. The allergy table also contains a field for medication ndc codes to map medication allergies to a specific medication.

Diagnosis Table

The diagnosis table contains the ICD9 Code to specifically identify the diagnosis, as well as an Acute indicator to flag acute instances of a diagnosis. Although there is a diagnosis description field, its contents are not standardized and thus not suitable for analysis. Instead, the ICD9 codes can be mapped to names from a table available online: List of ICD-9 Codes.

Medication Table

The medication table contains the NdcCode field, which specifically identifies medications, as well as the Medication Name.

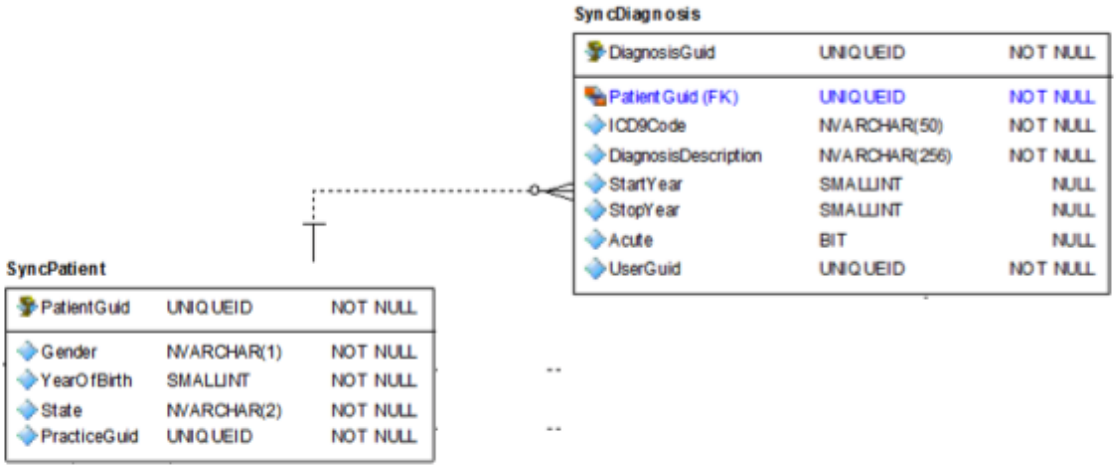


Figure 1: Patient and Diagnosis Tables

Transcript Table

The transcript table contains fields for height, weight, BMI, SystolicBP, DiastolicBP, respiratory rate, heart rate, and temperature.

Lab Observation Table

The lab observation table contains values from a variety of labs identified by an HL7 identifier. The text for these is more standardized and can be used to identify the type of lab the observations are for. The observation value and units fields provide the actual measurement values. There is an identifier for abnormal values as well as an abnormal flags field which indicates whether the abnormal value is high or low and to what degree.

Other Tables

There was insufficient data in the Immunization and Smoking tables for meaningful analysis. For the sake of simplicity, medication analysis was limited to the type of medication, so the prescription data was not analyzed. The condition table shown in the ER diagram was not included in the dataset.

Data Cleaning

Prior to analysis, the data included in the dataset required significant cleaning. The steps taken to clean the data are detailed below.

The raw data was contained in Comma Separated Values (CSV) files, one for each table. Each file was loaded into a separate data frame using read_csv(). Relevant fields were selected using select(). The example below shows the patient and diagnosis table files being read into data frames.

```
# read patient table into data frame.
patient <- read_csv("db/training_patient.csv") %>%
  select(-PracticeGuid)
```

```
# read diagnosis table into data frame.
diagnosis <- read_csv("db/training_diagnosis.csv") %>%
  select(DiagnosisGuid, DiagnosisDescription, PatientGuid, ICD9Code, StartYear, StopYear, Acute)
```

Since data was spread across different tables, the corresponding data frames needed to be joined in order to pull the data together into a single data frame.

```
# allergy data frame
patientAllergy <- inner_join(
  patient %>% select(PatientGuid, dmIndicator, diabetesStatus),
  allergy
)

# diagnosis data frame
patientDiagnosis <- inner_join(
  patient %>% select(PatientGuid, dmIndicator, diabetesStatus),
  diagnosis
)
```

Some of the data had too much differentiation and needed to be chunked in order to be analyzed. For example, the diagnoses were chunked into categories based on ICD-9 Code ranges.

```
# create diagCat column in patientDiagnosis containing diagnosis categories corresponding to ranges of
diagnosis$diagCat <-
  ifelse((as.integer(diagnosis$ICD9Code) < 140),
    "Infectious/Parasitic",
    ifelse((as.integer(diagnosis$ICD9Code) >= 140 &
      as.integer(diagnosis$ICD9Code) < 240),
      "Neoplasms",
      ifelse((as.integer(diagnosis$ICD9Code) >=240 &
        as.integer(diagnosis$ICD9Code) < 280),
          "Endocrine/Nutritional/Metabolic",
          ...
```

Another instance where there was simply too much data to analyze is for medication allergies. Since there is no readily available resource to map the codes into categories of medicine, a data frame was created that maps the code to the medication name for display purposes.

```
# create medicationMap data frame linking medication names to their NDC Codes
medicationMap <- medication %>% select(MedicationNdcCode, MedicationName) %>%
  group_by(MedicationNdcCode) %>% distinct(MedicationName) %>%
  arrange(MedicationNdcCode)
```

The top medication allergies were identified by sorting them by number of instances and filtering out the medications for which fewer than 20 diabetic patients have allergies.

```
# identify the medications for which at least 20 patients have allergies
topAllergyNdcCodes <- diabeticAllergyMeds %>%
  group_by(AllergyMedicationNdcCode) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  filter(n >= 20) %>%
  select(AllergyMedicationNdcCode)
```

Some data needed to be transformed into appropriate data types, filtered to remove outliers and apparent data errors (impossible values, etc.), and derived. The following example transforms several data fields in the transcript table into numeric values, filters out NA and negative values, and derives the pulse pressure from

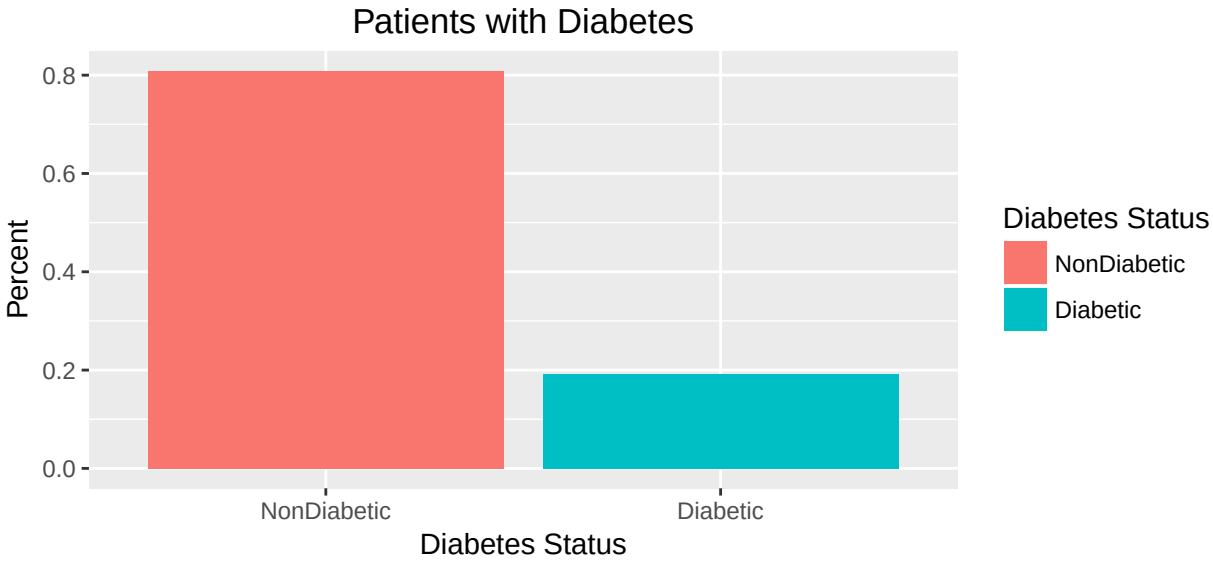
the systolic and diastolic values.

```
# change transcript Height & Weight to numeric types
transcript <- transform(transcript, Height = as.numeric(Height),
                Weight = as.numeric(Weight), Temperature = as.numeric(Temperature),
                RespiratoryRate = as.numeric(RespiratoryRate), HeartRate = as.numeric(HeartRate),
                SystolicBP = as.numeric(SystolicBP), DiastolicBP = as.numeric(DiastolicBP))

# add pulsePressure column to transcript (SystolicBP - DiastolicBP)
transcript <- transcript %>%
  filter(!is.na(SystolicBP)) %>% filter(!is.na(DiastolicBP)) %>%
  filter(SystolicBP > 0 & DiastolicBP > 0) %>%
  mutate(pulsePressure = SystolicBP - DiastolicBP)
```

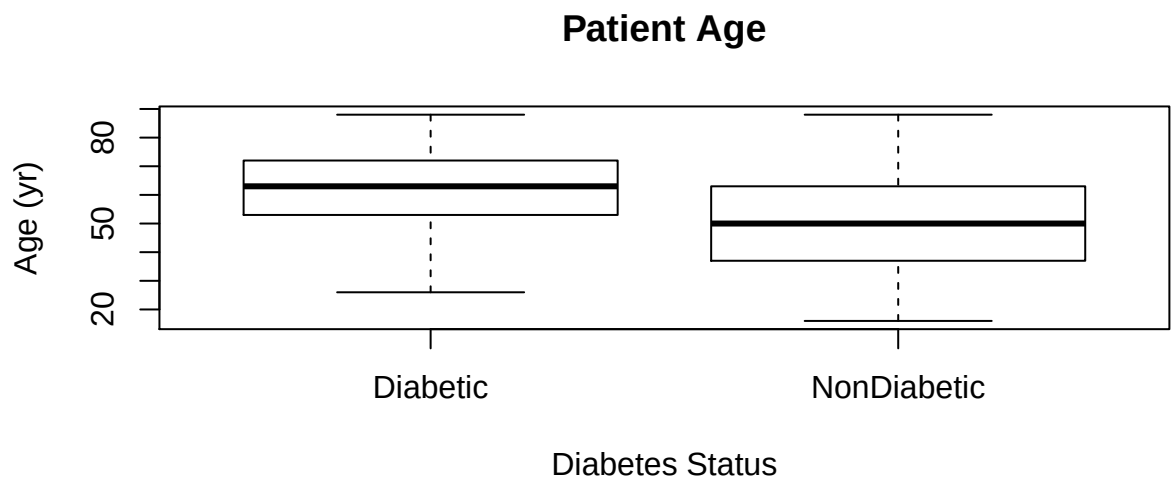
Statistical Evaluation of the EMR Dataset

Once the data was imported and cleaned, initial exploratory analysis could be performed. To start with, as a baseline comparison between the diabetic and non-diabetic patient populations in the dataset, 81% of the overall population is identified as non-diabetic and 19% is identified as diabetic.

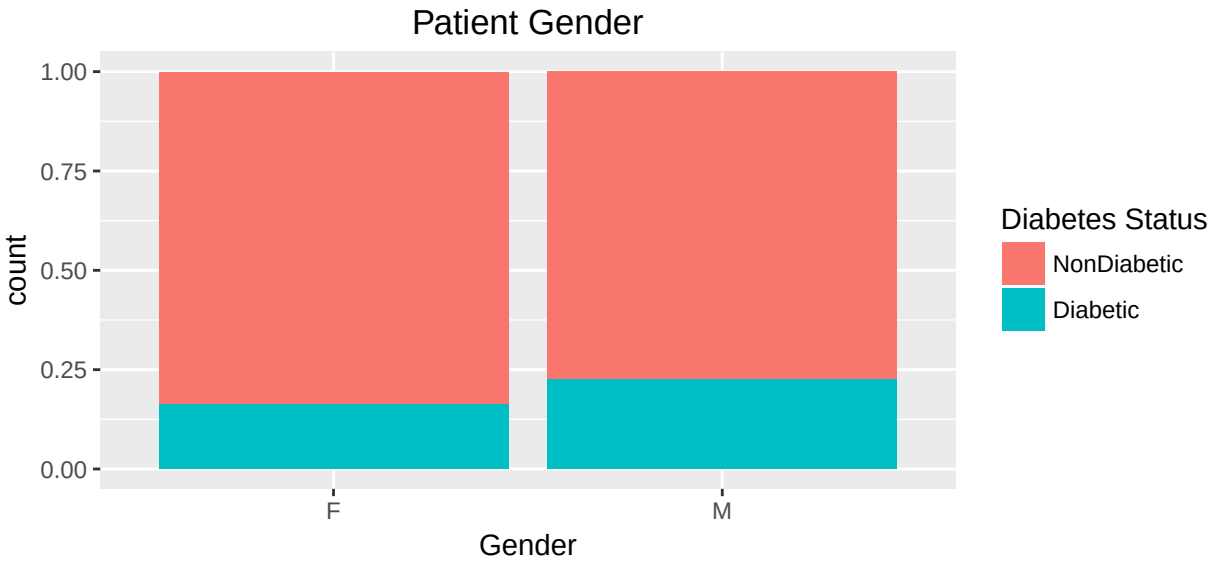


Age & Gender

It appears that the central tendency for Age is higher in the diabetic population than in the non-diabetic population.

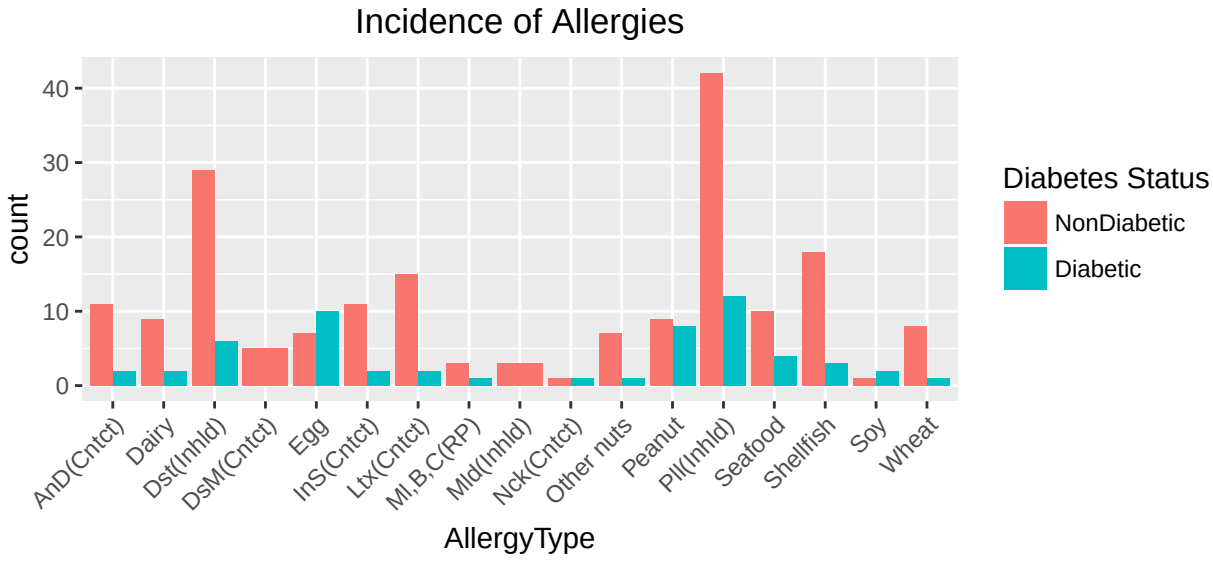


The number of patients with diabetes is similar between males and females, but the ratio of males with diabetes is higher.

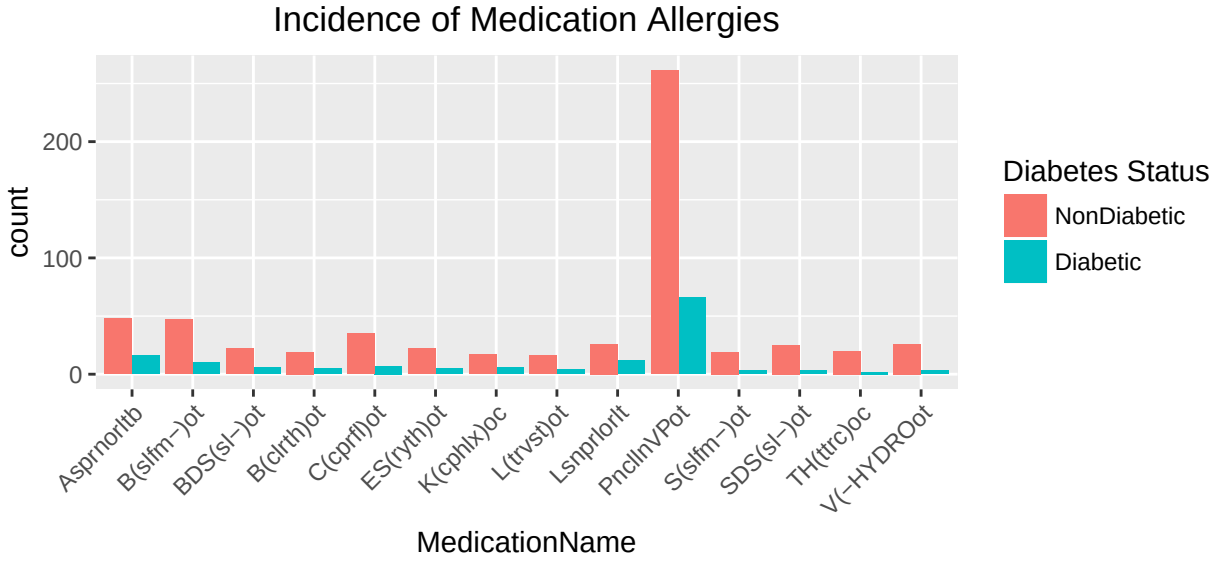


Allergies

Among non-medical allergy types, there appears to be a large proportion of diabetic patients with egg and peanut allergies in comparison to the general population. The number of diabetic patients with egg allergies actually outnumbers non-diabetic patients.

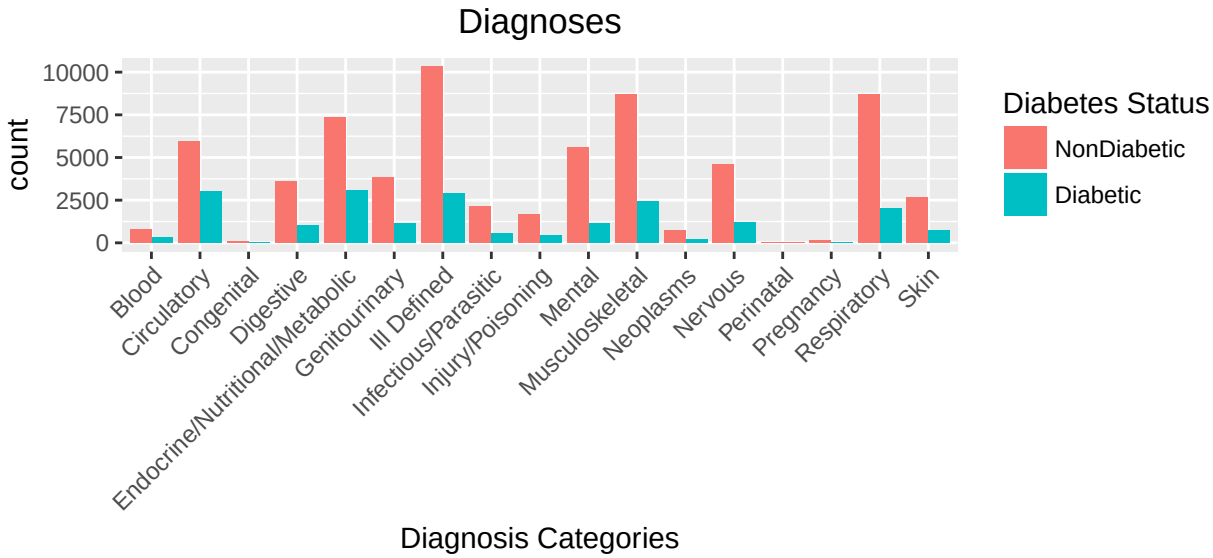


Among the allergies to medicine, the largest proportion of diabetic patients with an allergy compared to non-diabetic patients is to Lisinopril. However, the ratio is not significantly greater than the baseline.



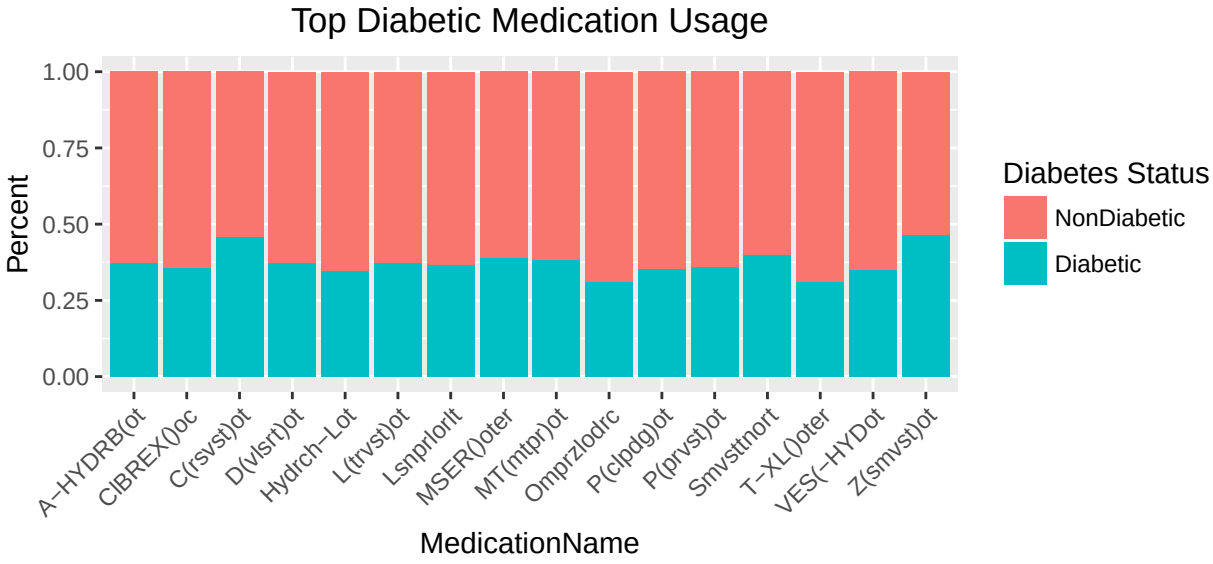
Diagnosis Categories

The diagnosis categories with the highest ratio between diabetic and non-diabetic patients are Circulatory and Endocrine/Nutritional/Metabolic.



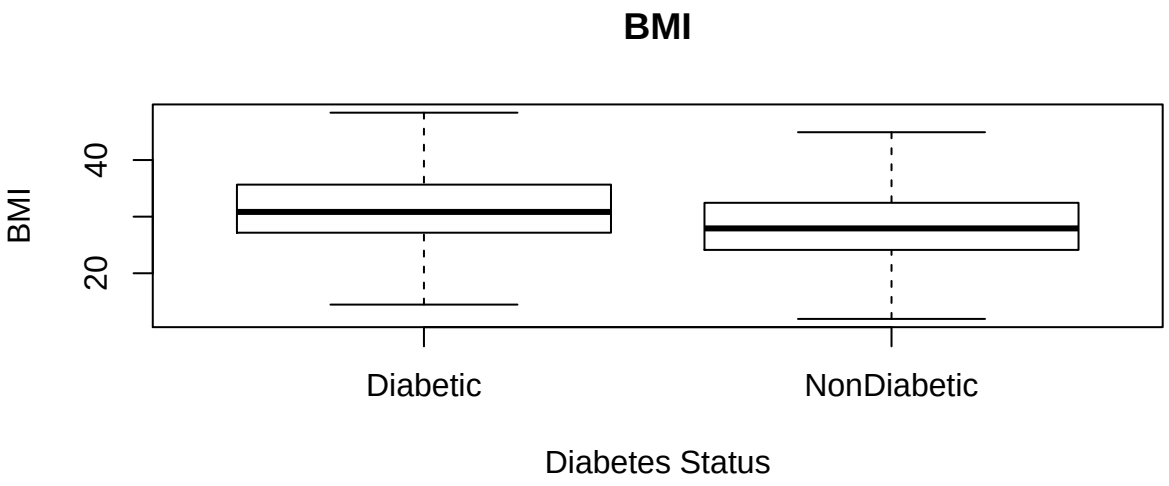
Prescription Data

There are 18 medications with at least 300 prescriptions and at least 30% use by diabetic patients.



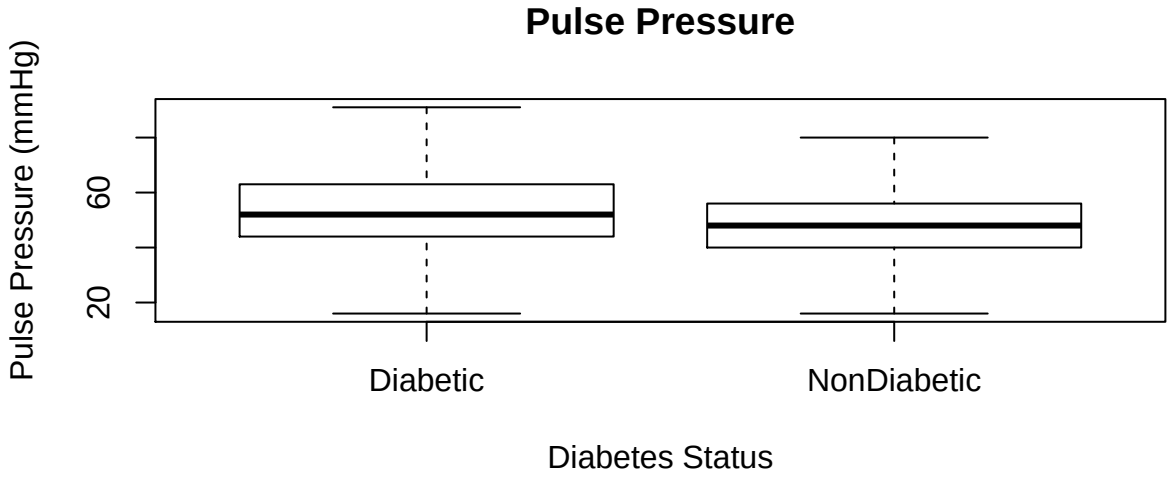
BMI

Diabetic patients have a slightly higher weight and shorter height than non-diabetic patients. Thus, their BMI trends higher.



Blood Pressure

The pulse pressure (Systolic BP - Diastolic BP) of diabetic patients is slightly higher than non-diabetic patients.



Lab Result Analysis

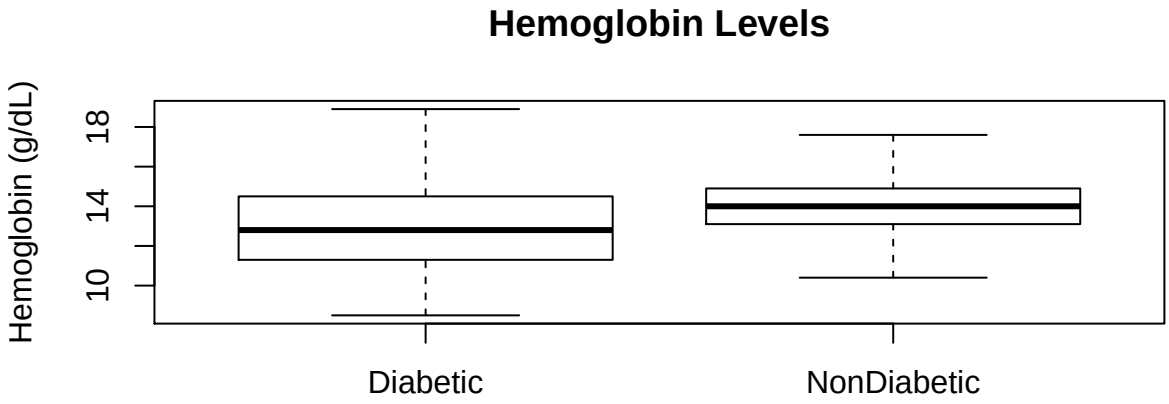
The lab results were limited to those for which there were sufficient abnormal readings data for the diabetic population. For each lab result, an overall measure of central tendency was analyzed, as well as an analysis of abnormal lab result status.

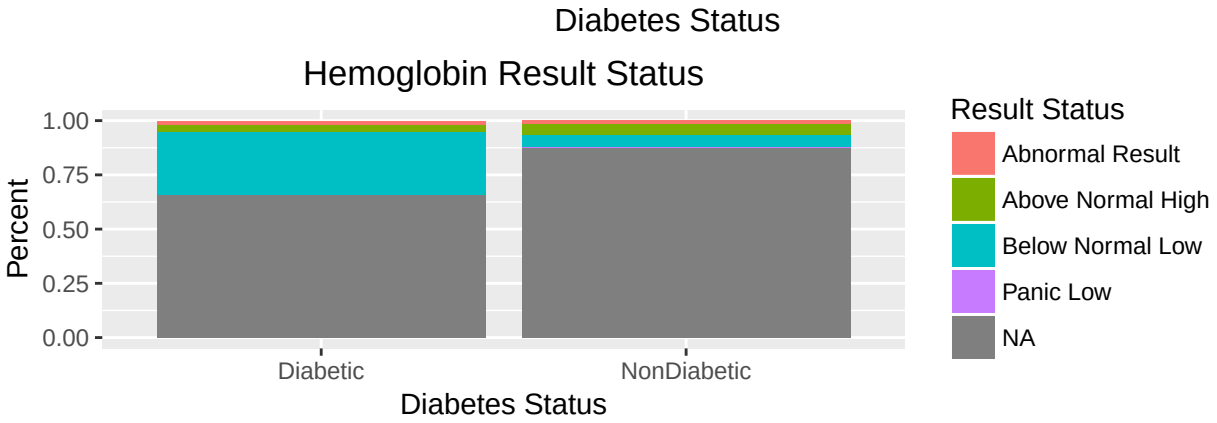
```
## # A tibble: 58 x 2
##    HL7Text                    n
##    <chr>                  <int>
##  1 Hemoglobin                39
##  2 Hematocrit                31
##  3 Chloride                  14
##  4 Triglyceride              14
##  5 Platelets                 13
##  6 RBC DISTRIBUTION WIDTH     9
##  7 ABS SEGS                   8
##  8 DIFFERENTIAL: SEGS         8
```

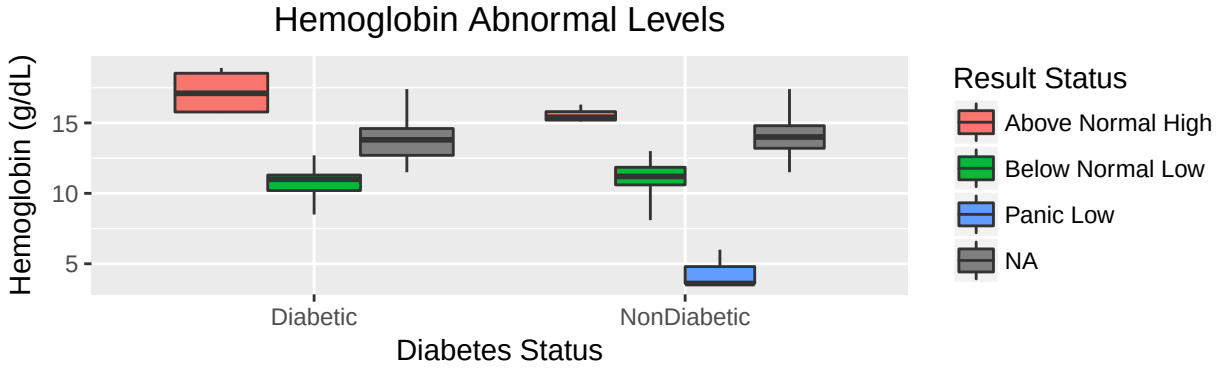
```
##  9 EOSINOPHILS                8
## 10 LYMPHOCYTES                8
## # ... with 48 more rows
```

Hemoglobin Lab Results

Hemoglobin levels have a lower central tendency in diabetic patients than in non-diabetic patients. When results recorded as abnormal are separated out, the diabetic population has a larger percentage of below normal readings. The central tendency of above normal readings were higher and of below normal readings were lower among the diabetic population. The central tendency of normal readings was slightly lower in the diabetic population.

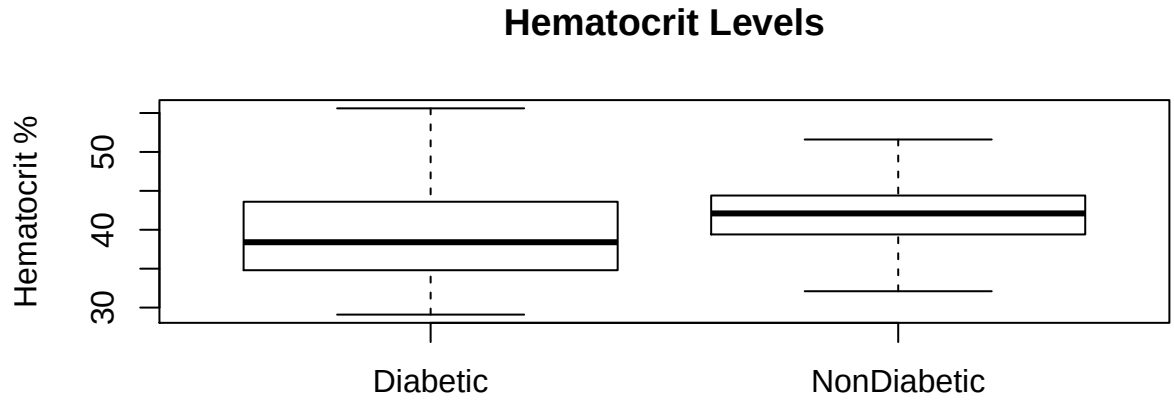


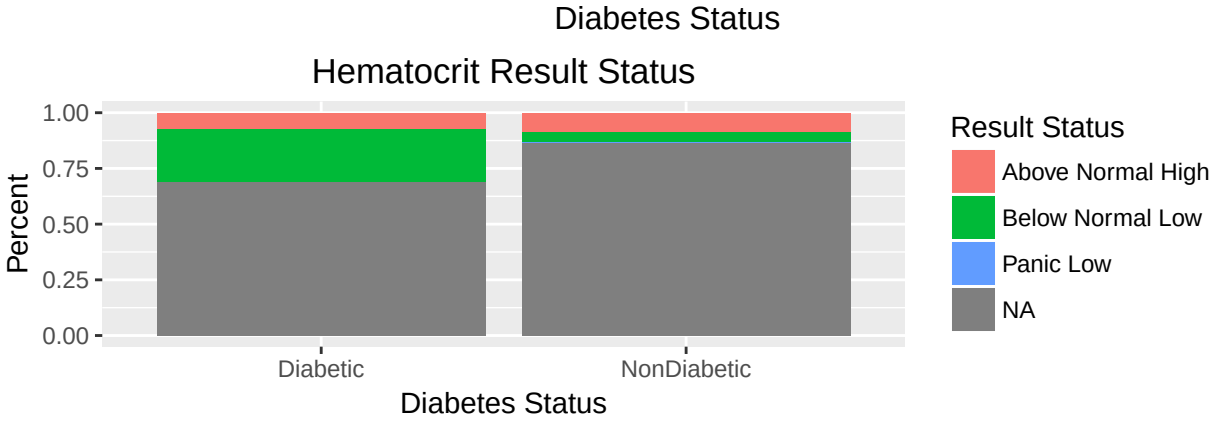


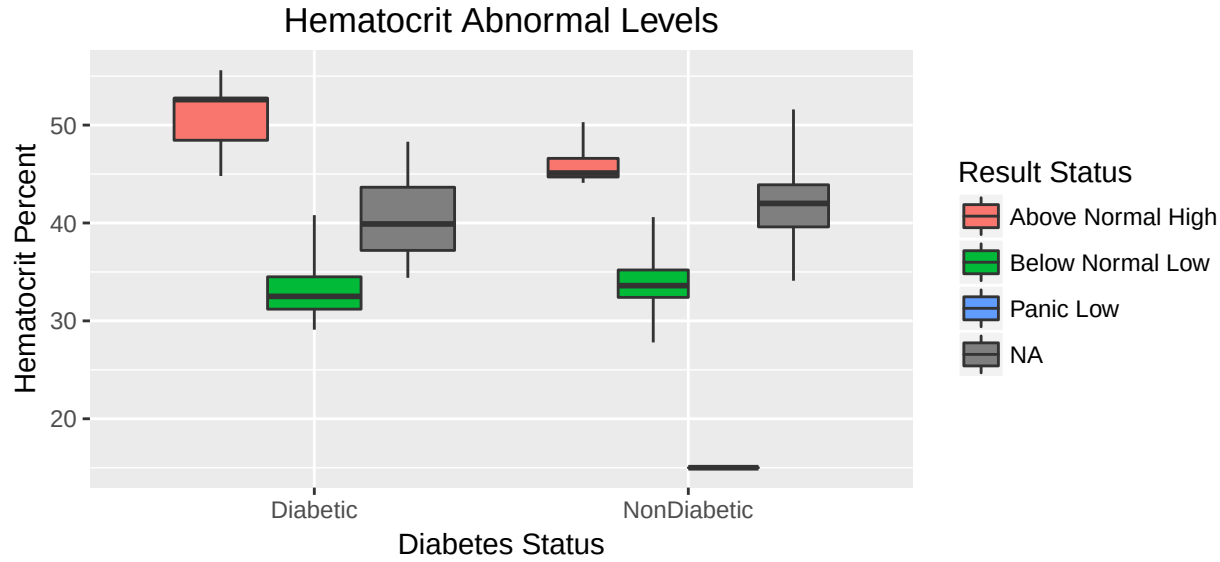


Hematocrit Lab Results

The central tendency of Hematocrit percentage was lower in the diabetic population than in the non-diabetic population. The ratio of above normal readings was lower and of below normal readings was higher among the diabetic population. The central tendency of above normal readings was higher and of below normal readings was lower in the diabetic population. The central tendency of normal readings was lower in the diabetic population.

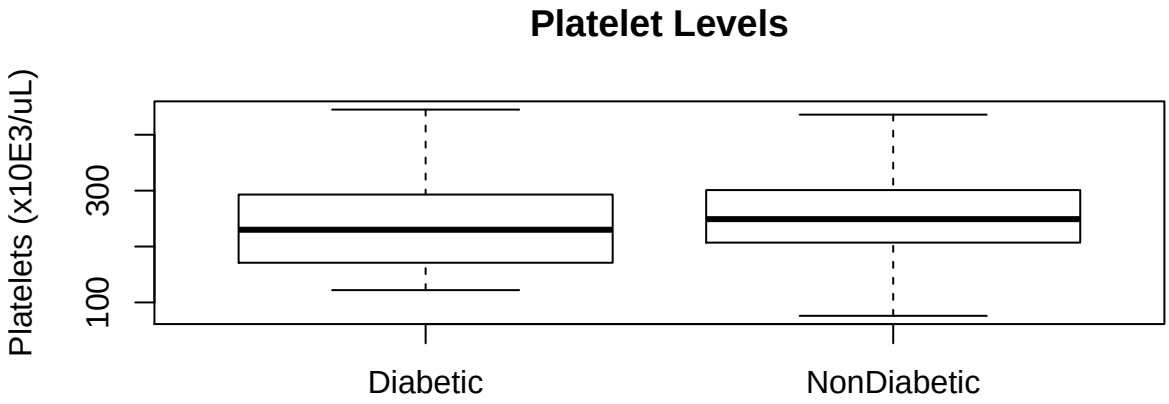


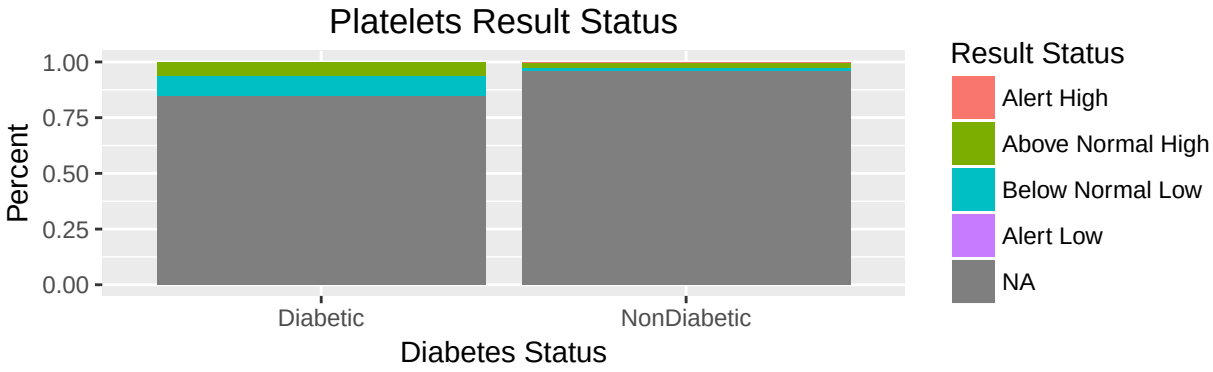


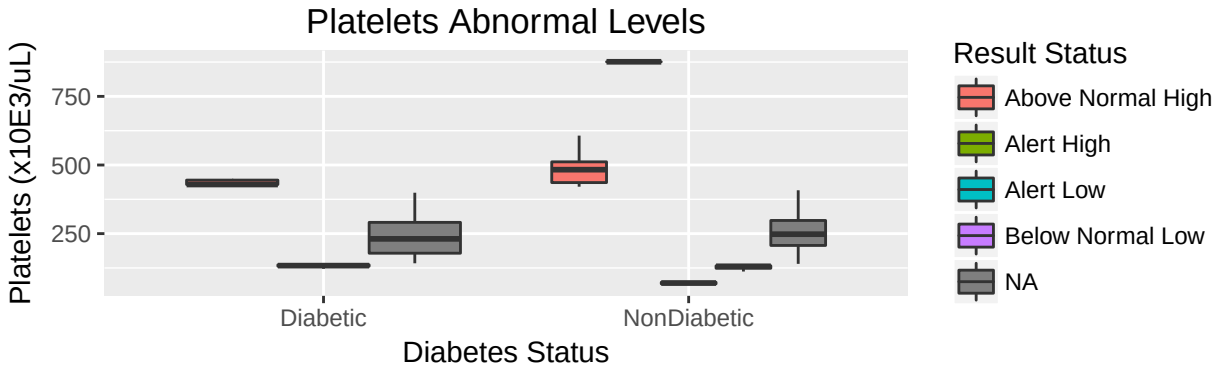


Platelets Lab Results

The central tendency of Platelet levels is lower in the diabetic population than in the non-diabetic population, particularly in the normal set. The ratio of both above and below normal readings are greater in the diabetic population, but the abnormal levels aren't as severe.

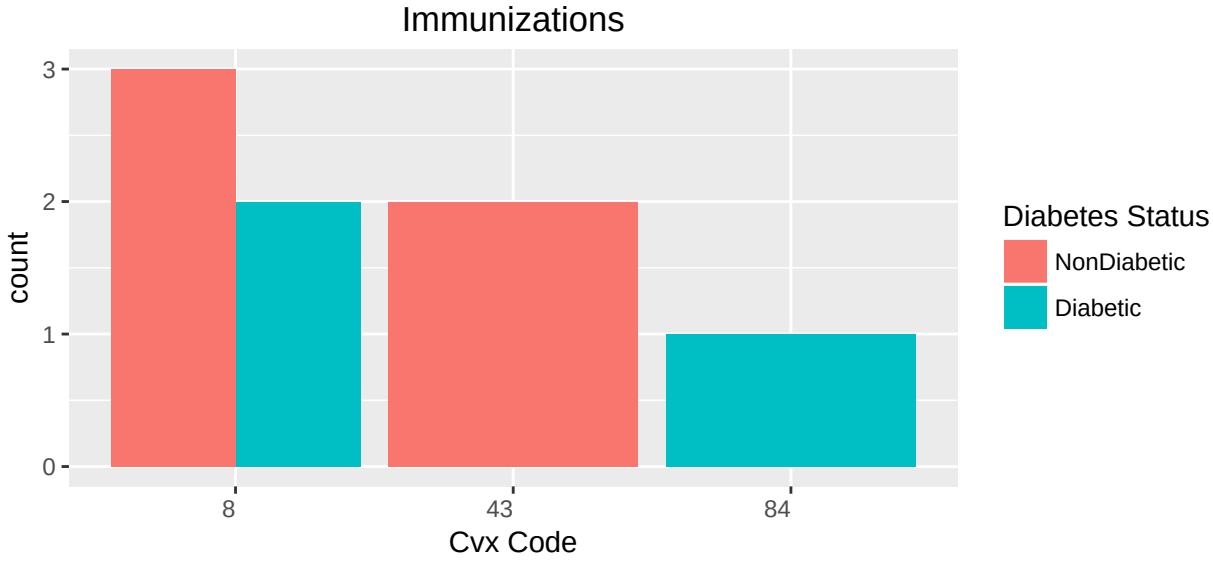






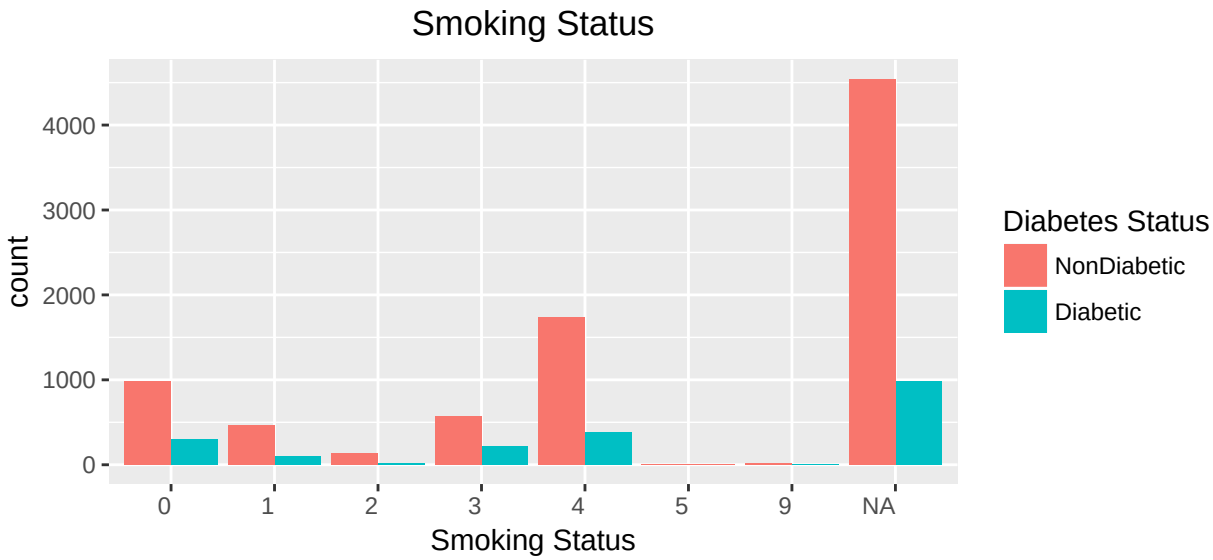
Immunizations

There were only 8 data points in the immunization table. This is not enough data to include in a predictive model.



Smoking Status

None of the smoking status data appears to be significantly different from the baseline.



Predictive Models

After some experimentation, Random Forest was selected as the type of predictive model to be used with the data set. For each model, data was split into training and test data subsets in order to evaluate the accuracy of the model. Also, cross validation was used in order to further evaluate the model.

Promising Variables for Data Models

After analyzing the data, the most promising variables that might contribute to a predictive model are:

- Age
- Gender
- Allergies
- Diagnosis Categories
- Prescriptions (?)
- BMI
- Pulse Pressure
- Hemoglobin Levels
- Hematocrit Levels
- Platelet Levels

Considerations for model construction

There are some details that need to be taken into account when constructing the data models. Some of the data is continuous and some of it is categorical. Each patient may have numerous observations, and some of them are easier to combine than others. For instance, a specific allergy might be linked to one transcript, which may contain a BMI and/or Pulse Pressure measurement, but others may not. All of the data can be joined with the transcript table, which determines which data can be condensed within a specific observation.

Although some lab observations are from the same lab panel and can thus be condensed into a single observation, none of the labs are associated with transcripts in the transcript table. Thus, all of the lab data is independent from the rest of the database.

Lab Results Data Predictive Model

The first predictive model was based on variables found in the lab observations. Because the lab observations do not connect with other data in the dataset, this model simply combined the desired lab data with the patient data (age and gender). In the exploratory analysis, the Hemoglobin, Hematocrit, and Platelet levels seemed most promising for a model. In order to use them in a model, we had to filter out the other lab observations, move the desired observations into their own columns, and then collapse them into a single row for each lab panel.

```
# platelet observations
platelets <- labObservation %>%
  filter(HL7Text == "Platelets" & !is.na(ObservationValue)) %>%
  select(LabPanelGuid, platelets = ObservationValue)

# hemoglobin observations
hemoglobin <- labObservation %>%
  filter(HL7Text == "Hemoglobin" & !is.na(ObservationValue)) %>%
  select(LabPanelGuid, hemoglobin = ObservationValue)

# hematocrit observations
```

```
hematocrit <- labObservation %>%
  filter(HL7Text == "Hematocrit" & !is.na(ObservationValue)) %>%
  select(LabPanelGuid, hematocrit = ObservationValue)
```

Some of the observation values were missing. Imputation was used to provide estimates for these values.

```
# Impute missing values
set.seed(1234)
labsImputed = complete(mice(labFrame))
```

After some experimentation, a random forest model based on age, gender, and lab results from hematocrit, hemoglobin, and platelet labs was created.

```
# Create random forest model
labForest = randomForest(
  dmIndicator ~ age + Gender + hematocrit + hemoglobin + platelets,
  data = labTrain, nodesize=25, ntree=500, na.action = na.omit, importance=TRUE)
```

According to a confusion matrix, the accuracy of the model is about 88%, though for a relatively small subset of the data for which the labs in question are available.

```
##    labPred
##       0   1
##   0 216   3
##   1  27   3
```

Cross validation indicates an accuracy of 89.6% and a Kappa of 31%.

```
## Random Forest
##
## 828 samples
##   5 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 745, 745, 745, 745, 745, 745, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.8961211  0.3146269
##   3     0.8888922  0.2932940
##   5     0.8901117  0.3053888
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

Diagnosis Data Predictive Model

Another subset of the data is the diagnosis data. This data was linked to the transcript data for various measurements taken on patient visits (blood pressure, etc.). But first, there were some apparent errors in the transcript data that were filtered out.

```
transSubset <- transcript %>% select (TranscriptGuid, PatientGuid, BMI, pulsePressure) %>%
  filter(BMI > 10 & BMI < 60 & pulsePressure > 20 & pulsePressure < 100)
```

In order to isolate the various diagnoses, the variable containing factors for the different diagnosis categories was split out into logical (TRUE/FALSE) columns and aggregated such that each patient has a single row containing all of the disease categories with which the patient has been diagnosed. The data was then split into training and test data frames.

```
diagAgg <- diagCat %>%
  group_by(PatientGuid, dmIndicator, BMI, age, Gender, pulsePressure) %>%
  summarise(
    Blood = as.logical(max(Blood)),
    Circulatory = as.logical(max(Circulatory)),
    Congenital = as.logical(max(Congenital)),
    Digestive = as.logical(max(Digestive)),
    Endo = as.logical(max(Endo)),
    ...
```

After some experimentation, a random forest model was created based on age, gender, BMI, pulse pressure, and the diagnosis categories of endocrine and circulatory diseases.

```
diagForest = randomForest(
  dmIndicator ~ age + Gender + Endo + Circulatory + pulsePressure + BMI,
  data = diagTrain, nodesize=25, ntree=500, na.action = na.omit, importance=TRUE)
```

According to a confusion matrix, the accuracy of this model is 79% across a much larger set of patient data.

```
##    diagPred
##        0    1
##   0 9637  317
##   1 2470  616
```

Cross validation indicated that the optimal tuning of this model produces an accuracy of 81.4% and a Kappa of 43%.

```
## Random Forest
##
## 43467 samples
##     6 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 39120, 39121, 39121, 39120, 39120, 39119, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.7894266  0.2356285
##   4     0.8139049  0.4348973
##   6     0.8136979  0.4390079
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 4.
```

Combined Model

A third model combined data from patients, allergies, diagnoses and transcript data into a single data frame, which was then split into training and test data sets. The large number of Medication categories exceeded

the limits of model functions, so only the identifiers in the prescription data were used to flatten the dataset as much as possible.

```
patientFrame <- left_join(
  left_join(
    left_join(
      left_join(
        patient %>% select(PatientGuid, dmIndicator, age, Gender) %>%
          transform(
            dmIndicator = as.factor(dmIndicator),
            Gender = as.factor(Gender)
          ),
        allergyFrame
      ),
      diagnosisFrame
    ),
    prescriptionFrame
  ),
  transcript %>% select(TranscriptGuid, BMI, pulsePressure)
)
```

A random forest model based on age, gender and allergy type and diagnosis category was trained. NA values were omitted. After some experimentation, the best results come from using Gender, age, and Allergy Type as independent variables.

```
set.seed(1234)
patientForest = randomForest(
  dmIndicator ~ Gender + age + AllergyType, data = patientTrain,
  nodesize=25, ntree=500, na.action = na.omit, importance=TRUE)
```

According to a confusion matrix, the accuracy of the model is about 95%. The subset of data included in the prediction is larger than the lab data predictive model but not as large as the diagnosis data predictive model.

```
##    predictForest
##       0   1
##   0 782   1
##   1  39  25
```

Cross validation an accuracy of 99% and a Kappa of 94.5%. This may be an indicator that the model is overfit and should be taken with a grain of salt.

```
## Random Forest
##
## 17627 samples
##     5 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 1154, 1155, 1154, 1155, 1155, 1155, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9898801  0.0000000
##   18    0.9984496  0.9459394
##   35    0.9984496  0.9459394
##
```

```
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 18.
```

Conclusion

Based on the preliminary results of both the exploratory analysis and predictive models, we can draw a few conclusions. Age and Gender are both good contributors to predictive models. Males are at a higher risk than females, as are older patients. Elevated hemoglobin and hematocrit levels and abnormal platelet levels are also predictive of Type 2 Diabetes. Pulse Pressure and BMI, as well as existence of other Endocrine or Circulatory diseases are also predictive of Type 2 Diabetes. There is some indication that Allergies may prove to be a strong predictor for Type 2 Diabetes, as well.

There are limitations to this study. There was difficulty in linking specific lab tests to specific doctor visits, so that it is difficult to get a thorough overall picture of the relationship between these results and other characteristics and readings from the patients. Also, due to the number of medications, they were not included in the models and their influence is therefore unknown. The unusually high accuracy of the combined model suggests that it may be overfit, and this may be due to the high number of factors in the categorical data elements.

Based on the findings of this project, it is recommended that further study be performed on the role of allergies in the risk of development of type 2 diabetes. Unlike factors such as BMI and blood pressure, which have well known connections to diabetes, this may yield some new understanding of risk factors. Further predictive models should be created with broader patient data sets to confirm and clarify these preliminary models.

References

1. National Diabetes Statistics Report 2017. National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation. [www.diabetes.org/assets/pdfs/basics/cdc-statistics-report-2017.pdf] Accessed 04/22/2018

2. Practice Fusion Diabetes Classification. Kaggle. [https://www.kaggle.com/c/pf2012-diabetes] Accessed 04/22/2018