Milestone Report

Keith Engwall

4/8/2018

Project Introduction

This project will attempt to create a predictive model for identifying patients with diabetes (or at least at risk for diabetes) using patient characteristics such as age, gender, allergies, comorbidities, prescriptions, and a variety of measures such as BMI, blood pressure, hemoglobin levels, etc.

Data

The project makes use of an EMR data set from 2012 of approx. 10,000 patients. The data set includes tables for patients, allergies, diagnoses, and prescriptions, as well as tables for transcripts of visit data and labs. In Figure 1 you can see the relationship diagram between the Patient and Diagnoisis tables. The full ER Diagram is available as Appendix A.

The diabetic patients are identified using a dmIndicator field in the patient table (not shown in the ER Diagram). The project will attempt to use patient data to predict which patients are in the diabetic group (dmIndicator = 1) as opposed to the non-diabetic group (dmIndicator = 0). Significant tables and fields containing the data this project will use 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

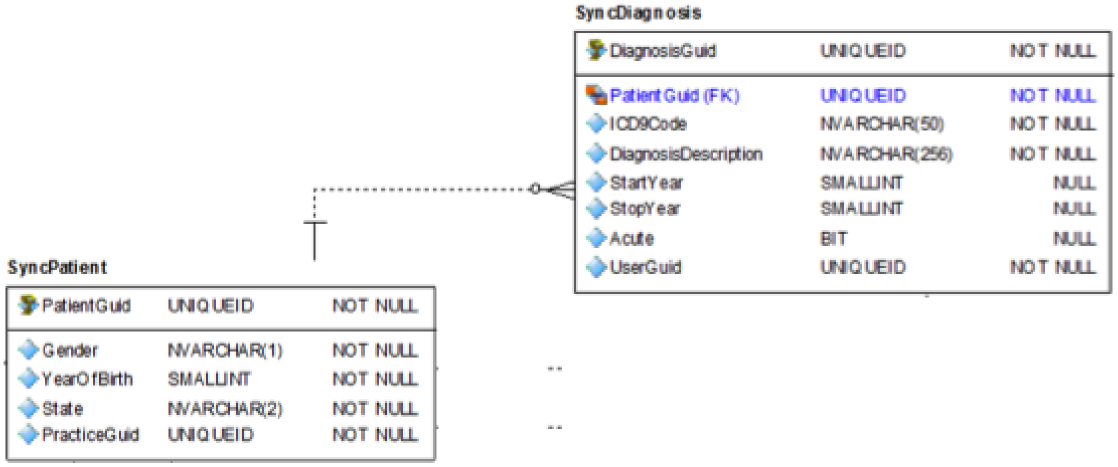
The allergy table contains a field for allergy type to identify the category of allergy, as well as for reaction name for the type of reaction and severity name for the severity of the allergic reaction. The allergy table also contains a field for medication ndc codes to map medication allergies to a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



SyncDiagnosis

DiagnosisGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
ICD9Code	NVARCHAR(50)	NOT NULL
DiagnosisDescription	NVARCHAR(256)	NOT NULL
StartYear	SMALLINT	NULL
StopYear	SMALLINT	NULL
Acute	BIT	NULL
UserGuid	UNIQUEID	NOT NULL

SyncPatient

PatientGuid	UNIQUEID	NOT NULL
Gender	NVARCHAR(1)	NOT NULL
YearOfBirth	SMALLINT	NOT NULL
State	NVARCHAR(2)	NOT NULL
PracticeGuid	UNIQUEID	NOT NULL

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.

Limitations

Although there is also an immunization table and a smoking status table, these do not provide sufficient data to analyze. Although the data structure allows for labs to link to visit transcripts, none do. Smoking data is too sparse and vague to use for analysis.

Data Cleaning

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.

```r
# 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. The example below shows the joining of patient, diagnosis, and transcript data.

```
# join the diagnosis frame and the transcript/diagnosis join frame, removing the join primary key (it's
diagJoin <- left_join(diagnosis, transDiag) %>% select(-TranscriptDiagnosisGuid)

# join the above frame with the transcript frame
diagTran <- left_join(diagJoin, transcript)

# join the patient frame with the above frame
patientDiagnosis <- left_join(patient, diagTran)
```

Some fields needed to be derived. For example, the age of the patients needed to be derived from the year of birth. The median year for the data set was 2010, and was used to derive the age from the patient data.

```
patient <- patient %>% mutate(age = 2010 - YearOfBirth)
```

The allergy field for the medication ndc code had to be renamed to disambiguate it from the prescription field of the same name. Numerous fields needed to be converted into numeric or integer types. In order to analyze systolic and diastolic blood pressure as a pair, a field for pulse pressure (systolic bp - diastolic bp) was added.

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

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. Some of the data needed to be filtered to remove insignificant data.

```
patientDiagnosis$diagCat <-
  ifelse((as.integer(patientDiagnosis$ICD9Code) < 140),
    "Infectious/Parasitic",
    ifelse((as.integer(patientDiagnosis$ICD9Code) >= 140 &
      as.integer(patientDiagnosis$ICD9Code) < 240),
      "Neoplasms",
      ifelse((as.integer(patientDiagnosis$ICD9Code) >=240 &
        as.integer(patientDiagnosis$ICD9Code) < 280),
          "Endocrine/Nutritional/Metabolic",
          ...
```

There were few if any records in the allergy table for the various types of medication allergies among the diabetic population. And for most of the remaining data, the ratio of diabetic patients with a particular medication allergy to non-diabetic patients was not of note. Therefore, the medication allergy data was filtered down to those for which there was at least one diabetic patient with an allergy and at least 20 patients overall with an allergy. A medication map needed to be created to map the NDC code for medication to the medication name.

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

# use inner join to filter patients to those
# with medication allergies, and pull in the names for the medications
```

```r
allergyMeds <- inner_join(patientAllergy,medicationMap, by = c("AllergyMedicationNdcCode" = "Medication

# identify the medications by name for which diabetic patients have allergies
diabeticMedNames <- allergyMeds %>% filter(dmIndicator == "1") %>% select(MedicationName) %>% distinct(

# use inner join to filter patients to those using the medications for which
# diabetic patients also have allergies (filter out all medication allergies
# for which diabetic patients do not have allergies)
diabeticAllergyMeds <- inner_join(allergyMeds,diabeticMedNames)

# 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)

# use inner join to filter data to those medications for which at least 20
# patients have allergies
diabeticAllergyMeds <- inner_join(allergyMeds,topAllergyNdcCodes)
```

For medication usage, the vast amount of data needed to be filtered still more. Only data where diabetic patients accounted for greater than 60% were included. Also, only data where there were greater than 300 records were included.

```r
# get a count of prescriptions for medications used by diabetic patients
diabeticMedicationList <- patientPrescription %>% filter(dmIndicator == 1) %>% group_by(MedicationName)

# join with a count of prescriptions for medications used by all patients
diabeticMedicationList <- inner_join(diabeticMedicationList, patientPrescription %>% group_by(Medicatio

# get the ratio between diabetic prescriptions and total prescriptions
diabeticMedicationList <- diabeticMedicationList %>%
  mutate(useRatio = n.x/n.y)

# had to tweak the filter to get a reasonably small set of the medications with the highest ratio of di
topDiabeticMedicationList <- diabeticMedicationList %>% filter(n.y > 300 & useRatio > .6) %>% arrange(d

# create a data frame limited to the top diabetic prescription list
topDiabeticPrescriptions <- inner_join(patientPrescription, topDiabeticMedicationList, by="MedicationNa
```

The abnormal flags needed to be reordered in order to display in a logical order (high to low) rather than in alphabetical order. Likewise, a field was added to provide a text equivalent for the diabetic patient indicator in order to be interpreted properly in the graphs.

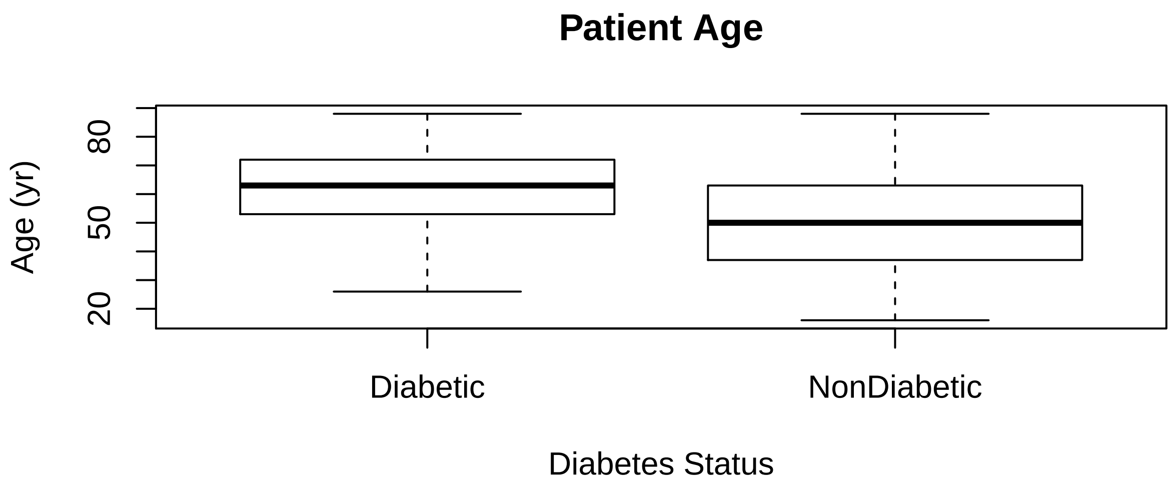
Initial Findings

In order to identify which data might be useful in predicting diabetic patients, some exploratory analysis was performed to compare the data between patients with diabetes and those without. The results of this analysis which indicate a potentially useful data field are shown below.

Age & Gender

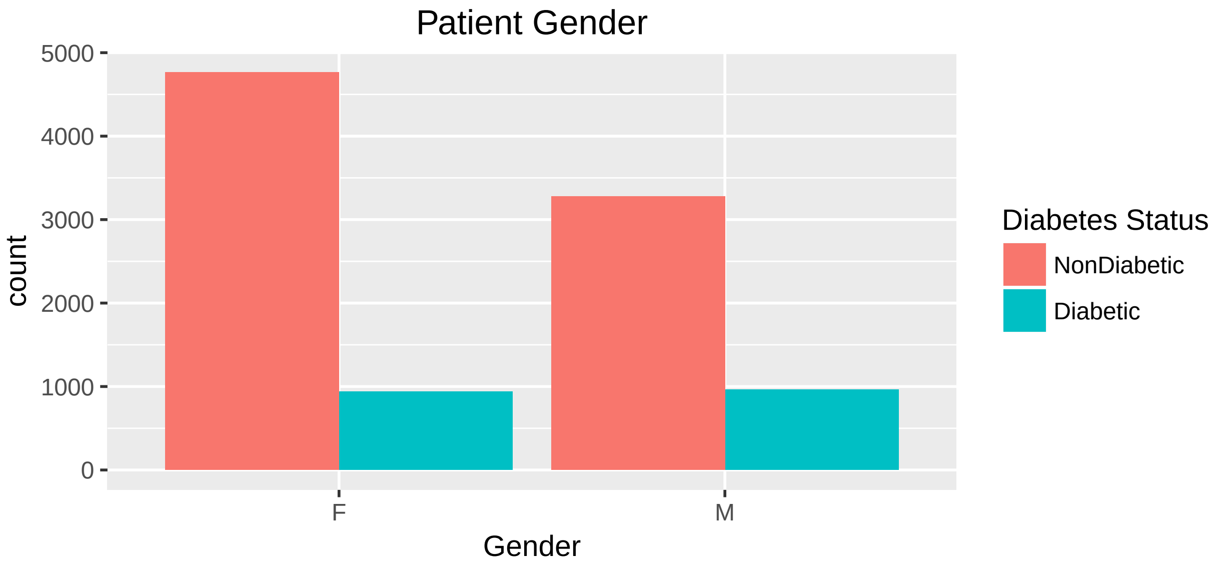
The diabetic population trends male and older than the non-diabetic population. Since Age data is continuous, a box plot was used.

```
boxplot(age~diabetesStatus,data=patient, outline = FALSE, main = "Patient Age", ylab = "Age (yr)", xlab
```



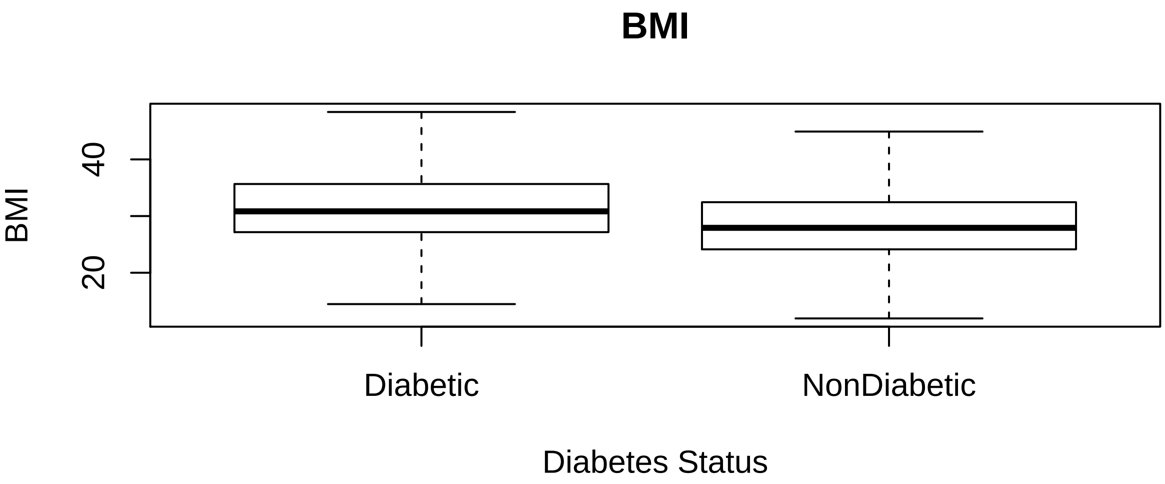
Since Gender data is categorical, a bar graph is used. Note that although the number of male and female diabetic patients are similar, the number of male non-diabetic patients is smaller, resulting in a larger ratio of male diabetic patients.

```
ggplot(arrange(patient, rev(dmIndicator)), aes(x=Gender,fill=factor(diabetesStatus, levels = c("NonDiab
  geom_bar(position = "dodge") +
  labs(title = "Patient Gender", fill="Diabetes Status") +
  theme(plot.title = element_text(hjust = "0.5"))
```



BMI

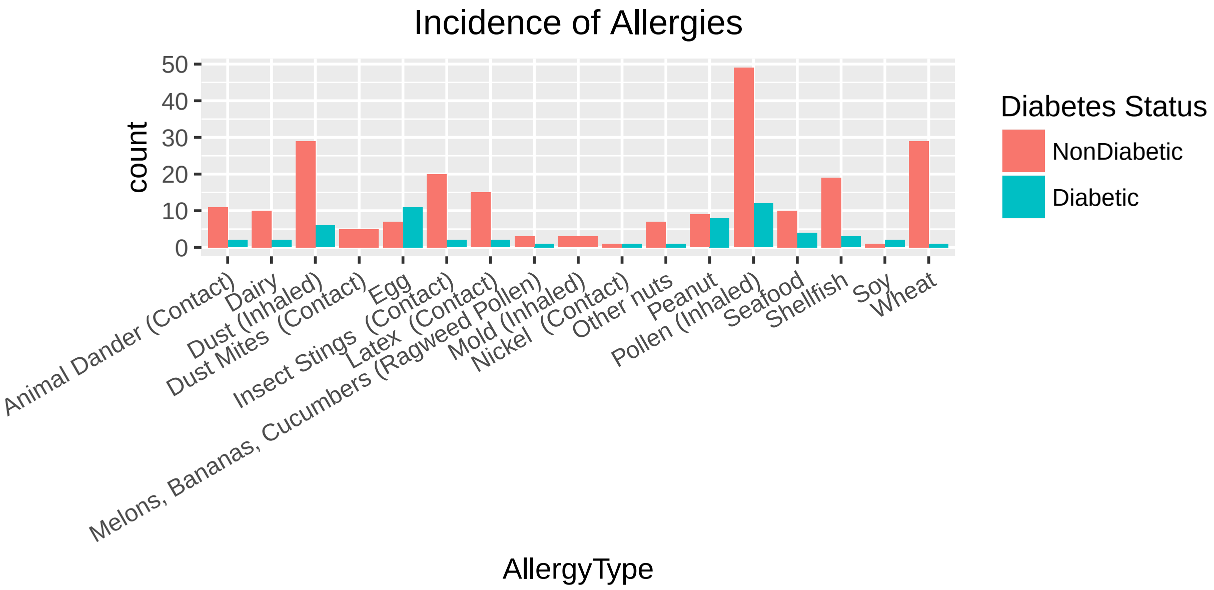
Diabetic patients have a slightly higher weight and shorter height than non-diabetic patients. Thus, their BMI trends higher.



Allergies

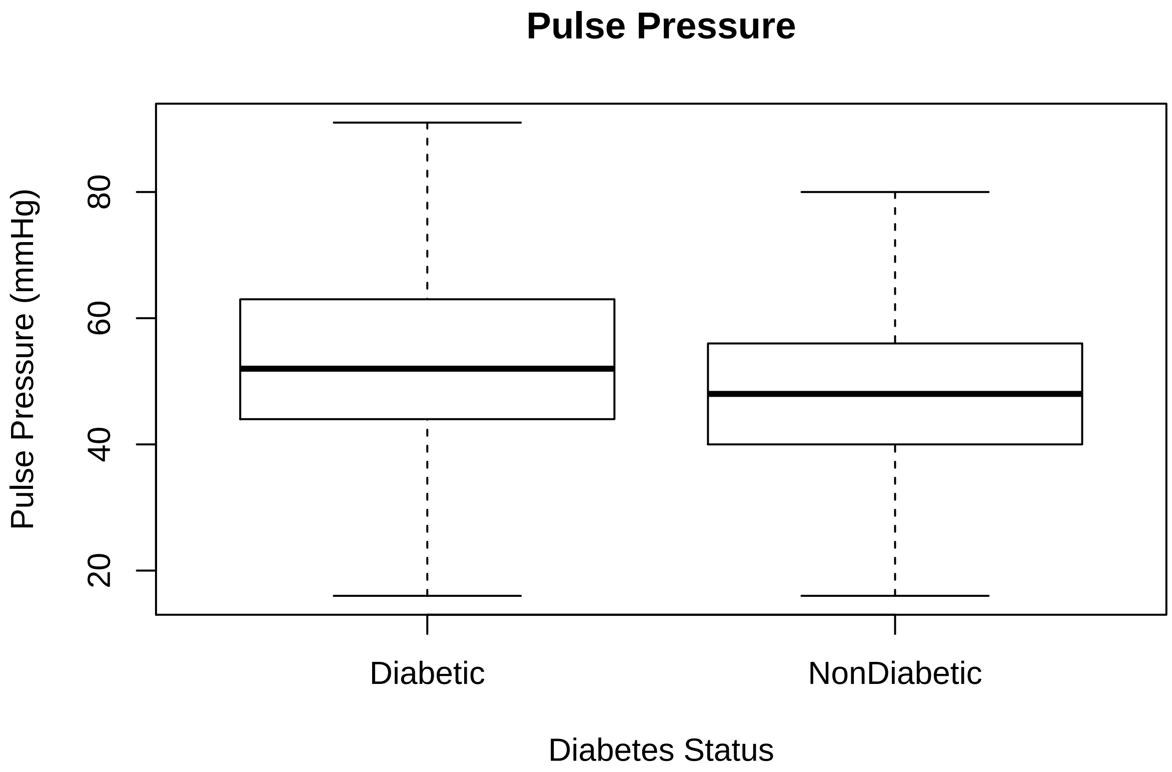
Among non-medical allergy types, there appears to be a large proportion of diabetic patients with egg and peanut allergies in comparison to the general population. The number of diabetic patients with egg allergies actually outnumbers non-diabetic patients. The plot required significant tweaking in order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

```
patientAllergy %>%
  filter(AllergyType != "Medication") %>%
  ggplot(aes(x=AllergyType, fill=factor(diabetesStatus, levels = c("NonDiabetic","Diabetic"))))+
  geom_bar(position="dodge") +
  labs(title = "Incidence of Allergies", fill = "Diabetes Status") +
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



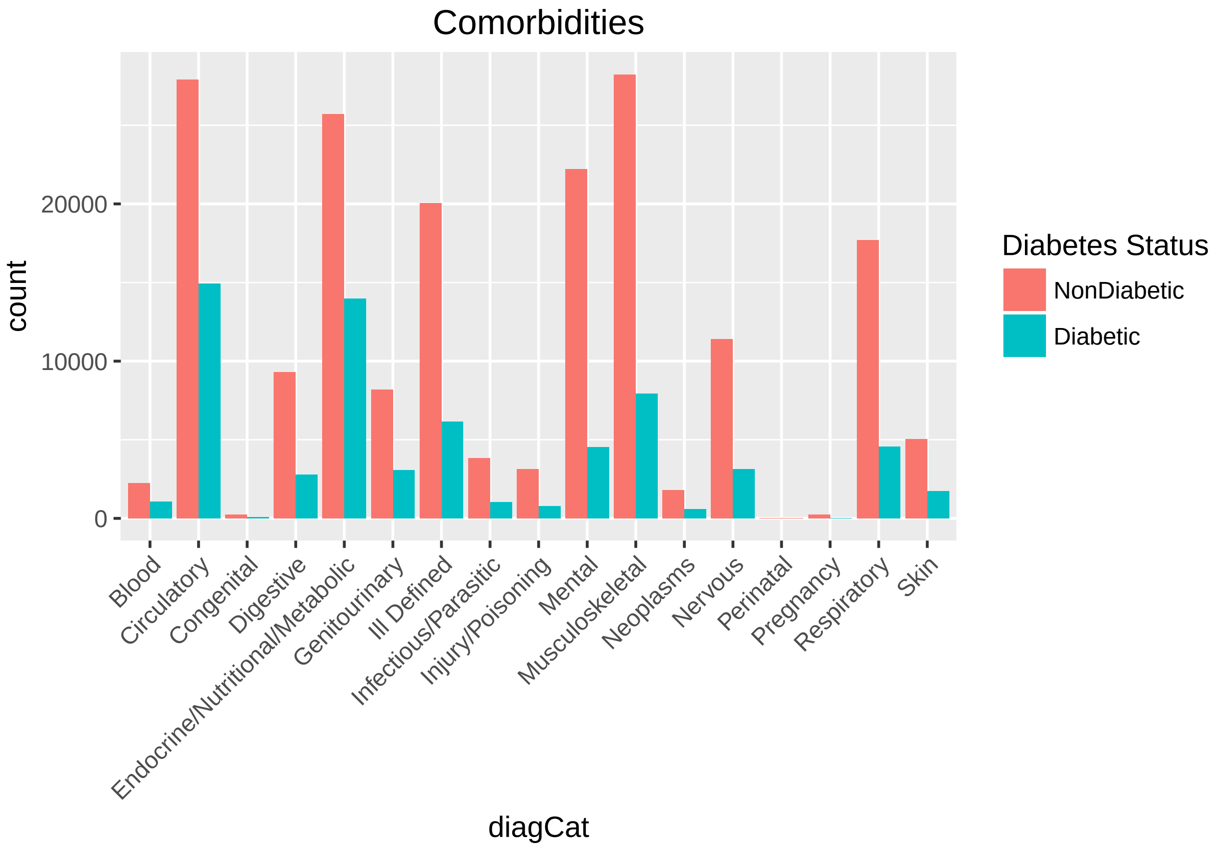
Blood Pressure

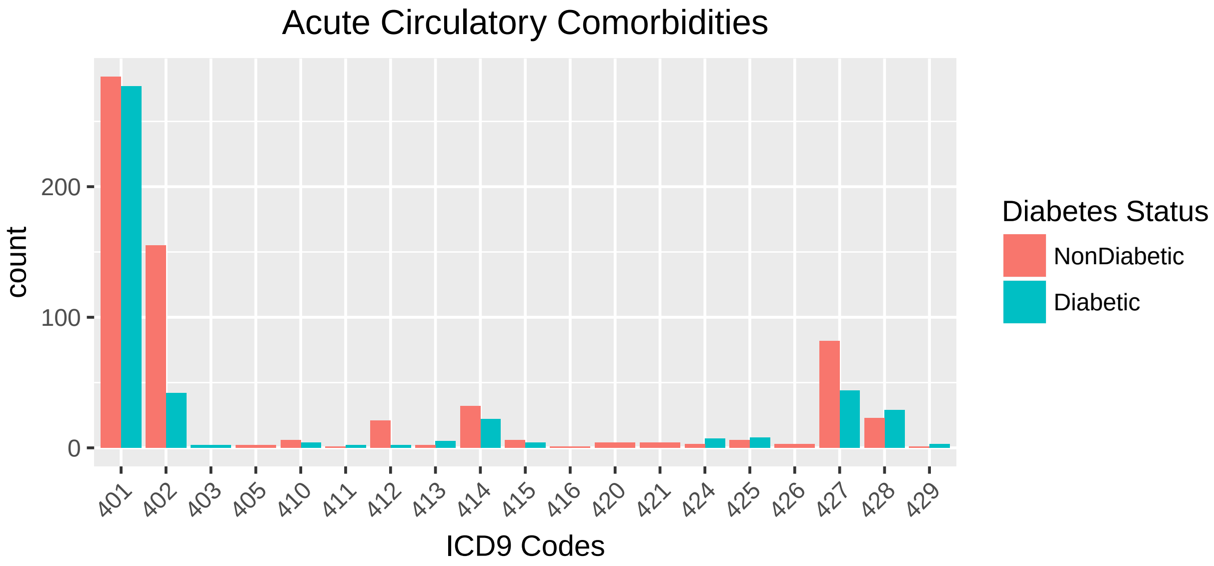
The pulse pressure (Systolic BP - Diastolic BP) of diabetic patients is slightly higher than non-diabetic patients.



Diagnosis Categories

To simplify analysis of diagnoses, the various diagnoses were divided into categories based on the ICD9 Codes. The diagnosis categories with the highest ratio between diabetic and non-diabetic patients is Circulatory. Within the Circulatory, there are almost as many diabetic patients with Essential Hypertension (ICD9 Code #401) as non-diabetic patients.



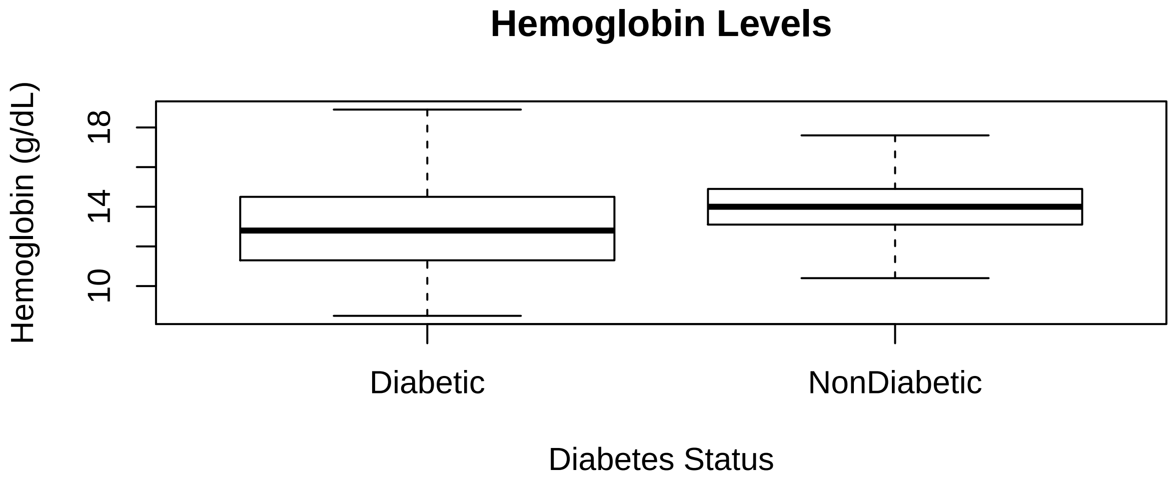


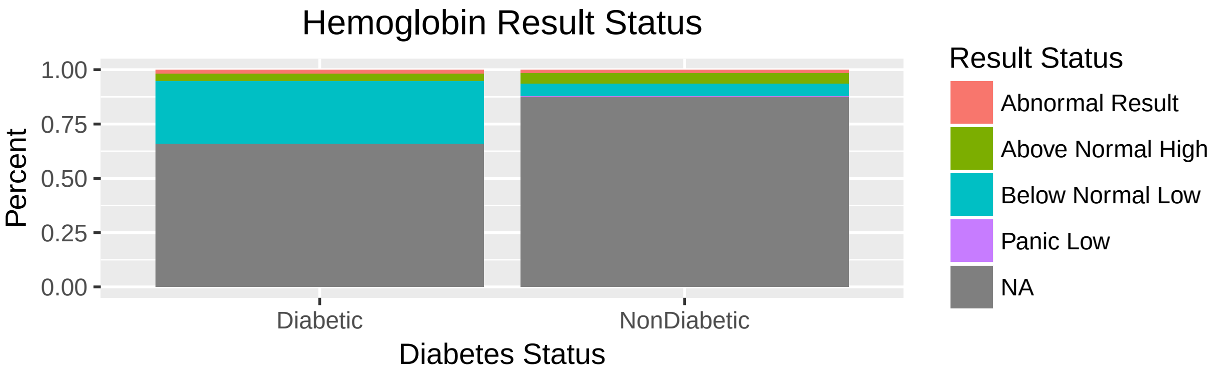
Lab Result Analysis

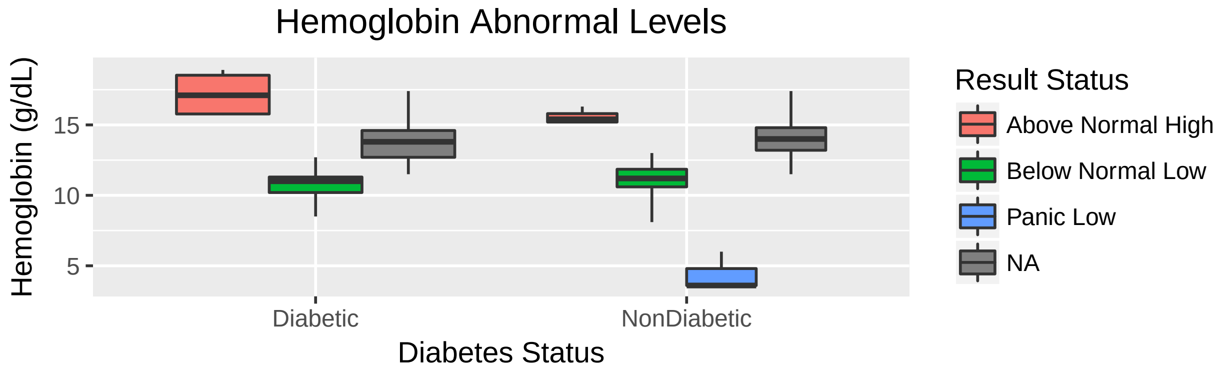
The lab results were limited to those for which there were sufficient abnormal readings data for the diabetic population. For each lab result, an overall measure of central tendency was analyzed, as well as an analysis of abnormal lab result status.

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.

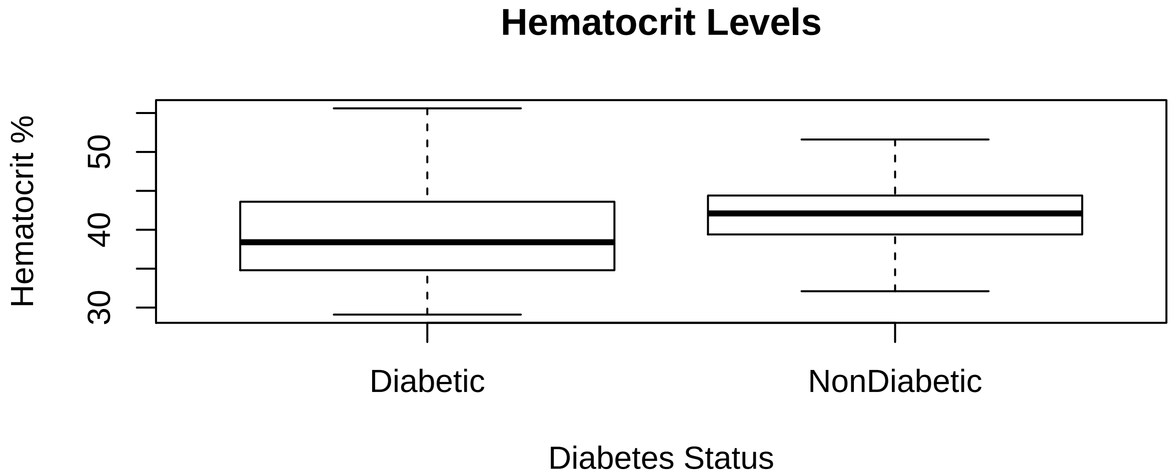


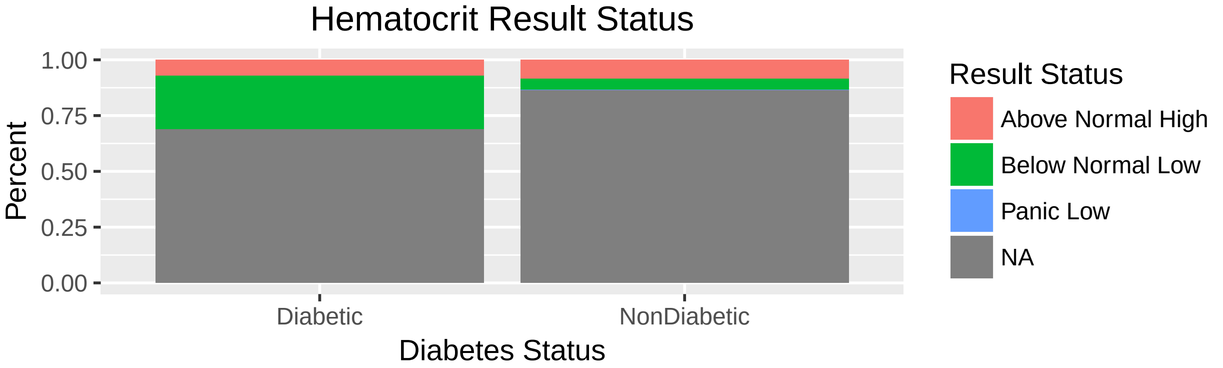


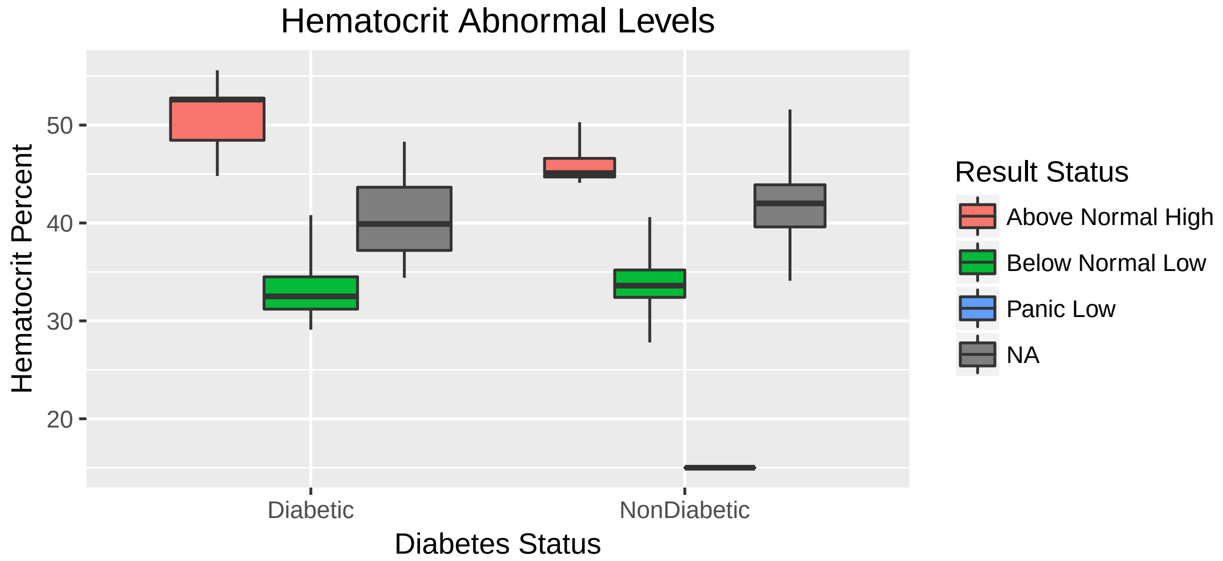


Hematocrit Lab Results

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.

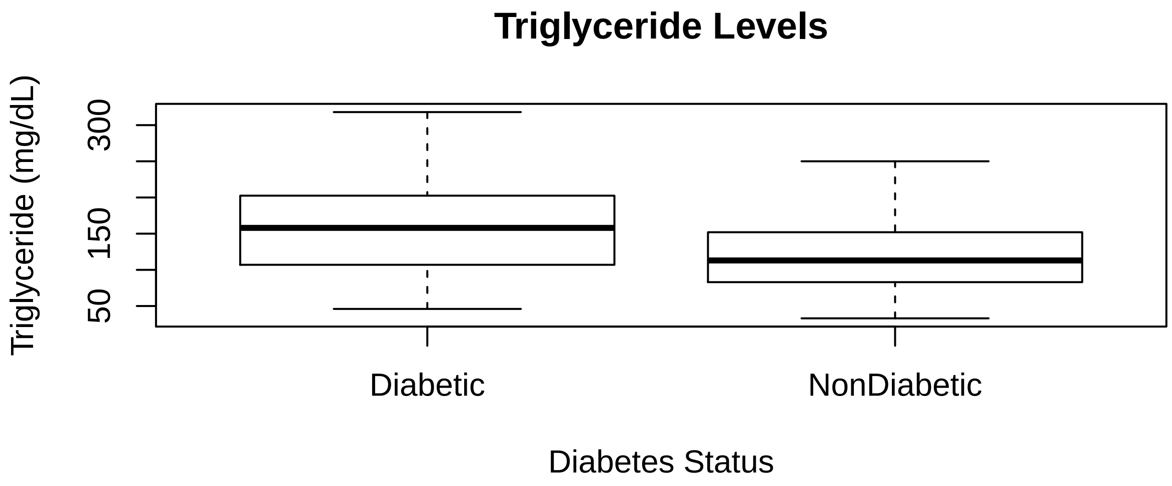


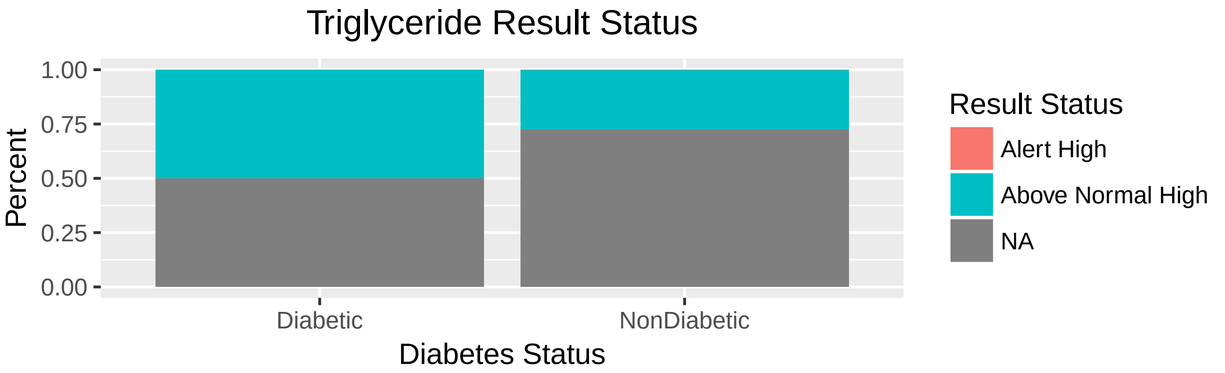


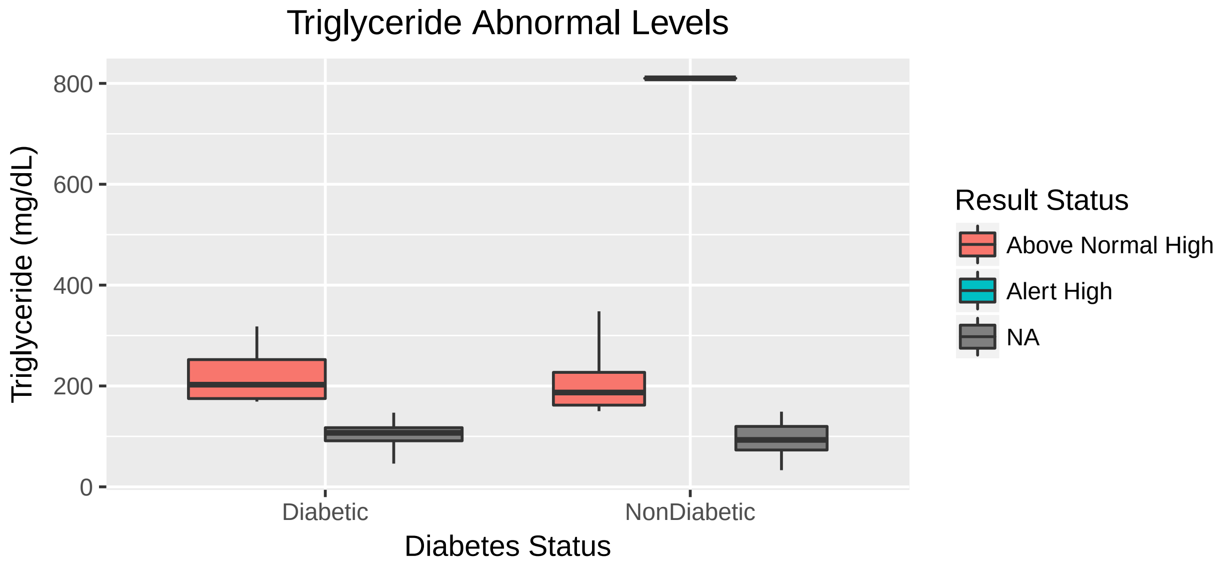


Triglyceride Lab Results

The central tendency of triglyceride levels is higher in the diabetic population than in the non-diabetic population. There is a higher ratio of above normal triglyceride readings in the diabetic population, and both the above normal and normal triglyceride levels are higher in the diabetic population.

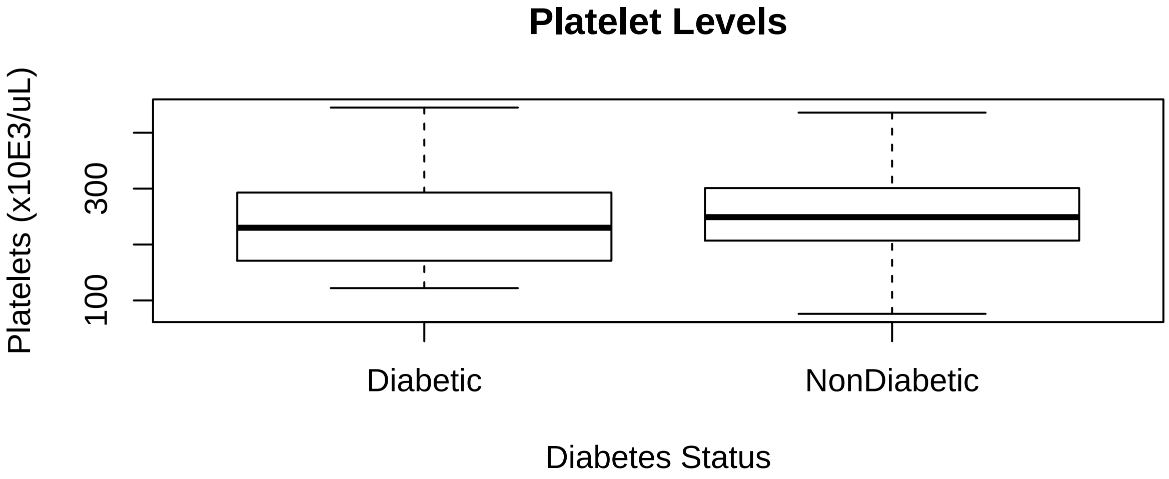


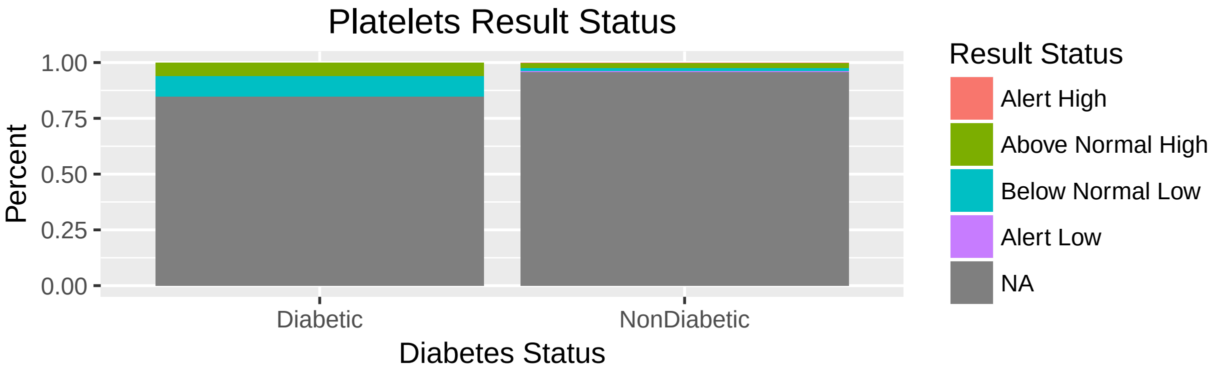


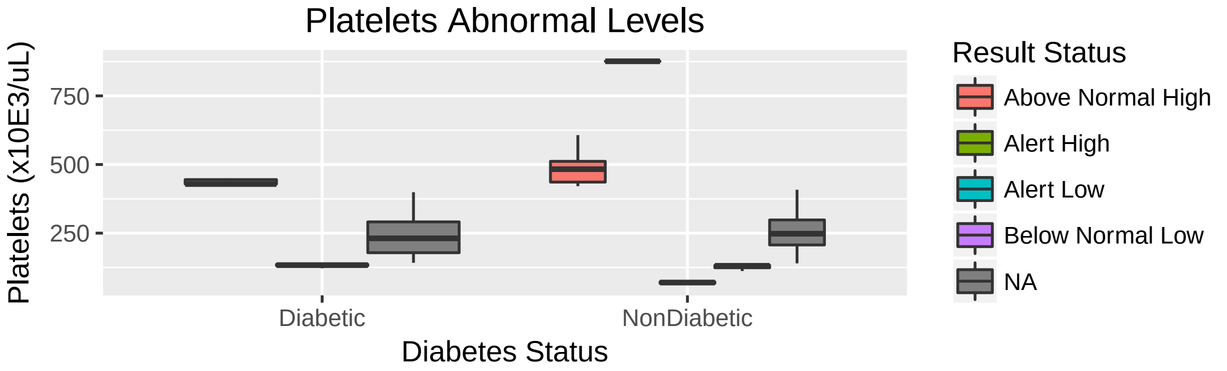


Platelets Lab Results

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.





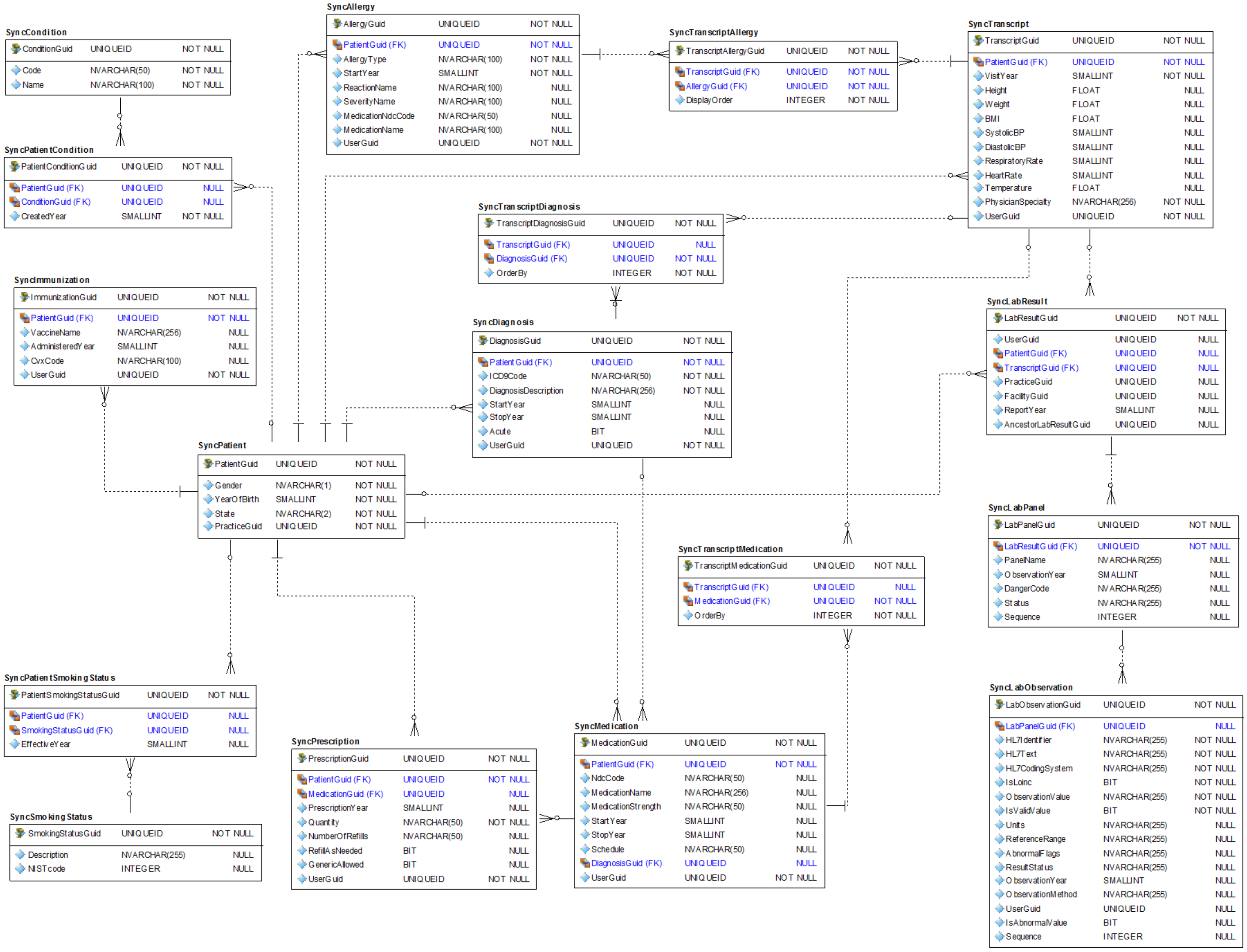


Approach

The intent is to use Random Forest, an extension of the Classification and Regression Trees (CART) predictive model. The data will need to be split into training and test data. The predictive quality of the model will be evaluated through cross validation. This model could be compared to a K-means cluster model using Gower distance for the mix of continuous and categorized data.

The dependent variable will be dmIndicator, and the independent variables will be some combination of Age, Gender, BMI, Allergies, Pulse Pressure, Diagnosis category, Hemoglobin levels, Hematocrit levels, Triglyceride levels, and Platelet levels. The final report will outline the process of how this model is developed, any further data cleaning required, and evaluation.

Appendix A: ER Diagram



SyncCondition

ConditionGuid	UNIQUEID	NOT NULL
Code	NVARCHAR(50)	NOT NULL
Name	NVARCHAR(100)	NOT NULL

SyncPatientCondition

PatientConditionGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NULL
ConditionGuid (FK)	UNIQUEID	NULL
CreatedYear	SMALLINT	NOT NULL

SyncAllergy

AllergyGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
AllergyType	NVARCHAR(100)	NOT NULL
StartYear	SMALLINT	NOT NULL
ReactionName	NVARCHAR(100)	NULL
SeverityName	NVARCHAR(100)	NULL
MedicationNdcCode	NVARCHAR(50)	NULL
MedicationName	NVARCHAR(100)	NULL
UserGuid	UNIQUEID	NOT NULL

SyncTranscriptAllergy

TranscriptAllergyGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NOT NULL
AllergyGuid (FK)	UNIQUEID	NOT NULL
DisplayOrder	INTEGER	NOT NULL

SyncTranscript

TranscriptGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
VisitYear	SMALLINT	NOT NULL
Height	FLOAT	NULL
Weight	FLOAT	NULL
BMI	FLOAT	NULL
SystolicBP	SMALLINT	NULL
DiastolicBP	SMALLINT	NULL
RespiratoryRate	SMALLINT	NULL
HeartRate	SMALLINT	NULL
Temperature	FLOAT	NULL
PhysicianSpecialty	NVARCHAR(256)	NOT NULL
UserGuid	UNIQUEID	NOT NULL

SyncImmunization

ImmunizationGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
VaccineName	NVARCHAR(256)	NULL
AdministeredYear	SMALLINT	NULL
CvxCode	NVARCHAR(100)	NULL
UserGuid	UNIQUEID	NOT NULL

SyncTranscriptDiagnosis

TranscriptDiagnosisGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NULL
DiagnosisGuid (FK)	UNIQUEID	NOT NULL
OrderBy	INTEGER	NOT NULL

SyncDiagnosis

DiagnosisGuid	UNIQUEID	NOT NULL
Patient Guid (FK)	UNIQUEID	NOT NULL
ICD9Code	NVARCHAR(50)	NOT NULL
DiagnosisDescription	NVARCHAR(256)	NOT NULL
StartYear	SMALLINT	NULL
StopYear	SMALLINT	NULL
Acute	BIT	NULL
UserGuid	UNIQUEID	NOT NULL

SyncLabResult

LabResultGuid	UNIQUEID	NOT NULL
UserGuid	UNIQUEID	NULL
PatientGuid (FK)	UNIQUEID	NULL
TranscriptGuid (FK)	UNIQUEID	NULL
PracticeGuid	UNIQUEID	NULL
FacilityGuid	UNIQUEID	NULL
ReportYear	SMALLINT	NULL
AncestorLabResultGuid	UNIQUEID	NULL

SyncPatient

PatientGuid	UNIQUEID	NOT NULL
Gender	NVARCHAR(1)	NOT NULL
YearOfBirth	SMALLINT	NOT NULL
State	NVARCHAR(2)	NOT NULL
PracticeGuid	UNIQUEID	NOT NULL

SyncLabPanel

LabPanelGuid	UNIQUEID	NOT NULL
LabResultGuid (FK)	UNIQUEID	NOT NULL
PanelName	NVARCHAR(255)	NULL
ObservationYear	SMALLINT	NULL
DangerCode	NVARCHAR(255)	NULL
Status	NVARCHAR(255)	NULL
Sequence	INTEGER	NULL

SyncTranscriptMedication

TranscriptMedicationGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NULL
MedicationGuid (FK)	UNIQUEID	NOT NULL
OrderBy	INTEGER	NOT NULL

SyncPatientSmokingStatus

PatientSmokingStatusGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NULL
SmokingStatusGuid (FK)	UNIQUEID	NULL
EffectiveYear	SMALLINT	NULL

SyncSmokingStatus

SmokingStatusGuid	UNIQUEID	NOT NULL
Description	NVARCHAR(255)	NULL
NISTcode	INTEGER	NULL

SyncPrescription

PrescriptionGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
MedicationGuid (FK)	UNIQUEID	NULL
PrescriptionYear	SMALLINT	NULL
Quantity	NVARCHAR(50)	NOT NULL
NumberOfRefills	NVARCHAR(50)	NULL
RefillAsNeeded	BIT	NULL
GenericAllowed	BIT	NULL
UserGuid	UNIQUEID	NOT NULL

SyncMedication

MedicationGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
NdcCode	NVARCHAR(50)	NULL
MedicationName	NVARCHAR(256)	NULL
MedicationStrength	NVARCHAR(50)	NULL
StartYear	SMALLINT	NULL
StopYear	SMALLINT	NULL
Schedule	NVARCHAR(50)	NULL
DiagnosisGuid (FK)	UNIQUEID	NULL
UserGuid	UNIQUEID	NOT NULL

SyncLabObservation

LabObservationGuid	UNIQUEID	NOT NULL
LabPanelGuid (FK)	UNIQUEID	NULL
HL7Identifier	NVARCHAR(255)	NOT NULL
HL7Text	NVARCHAR(255)	NOT NULL
HL7CodingSystem	NVARCHAR(255)	NOT NULL
IsLoinc	BIT	NOT NULL
ObservationValue	NVARCHAR(255)	NOT NULL
IsValidValue	BIT	NOT NULL
Units	NVARCHAR(255)	NULL
ReferenceRange	NVARCHAR(255)	NULL
AbnormalFlags	NVARCHAR(255)	NULL
ResultStatus	NVARCHAR(255)	NULL
ObservationYear	SMALLINT	NULL
ObservationMethod	NVARCHAR(255)	NULL
UserGuid	UNIQUEID	NULL
IsAbnormalValue	BIT	NULL
Sequence	INTEGER	NULL

© 2012 Practice Fusion