EMR Predictive Models for Patients with Diabetes

Keith Engwall

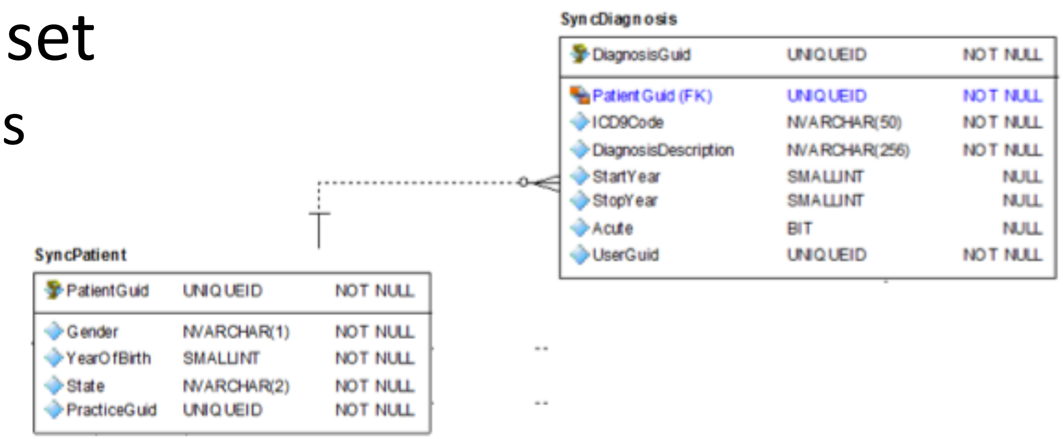
4/30/2018

Introduction

- Type 2 Diabetes accounts for 90-95% of all diabetes cases among adults in United States (2015 data)
  - 23 million diagnosed cases
  - 7.2 million undiagnosed cases[1]
- Predictive models may be used to identify undiagnosed individuals
- This project develops and evaluates predictive models for identifying patients with diabetes using an Electronic Medical Record (EMR) dataset

Data

- 2012 EMR data set
  - 10,000 patients
  - SQL tables
    - Patients
    - Allergies
    - Diagnoses
    - Prescriptions
    - Transcripts of visit data
    - Labs
  - Target patients identified with *dmIndicator* field

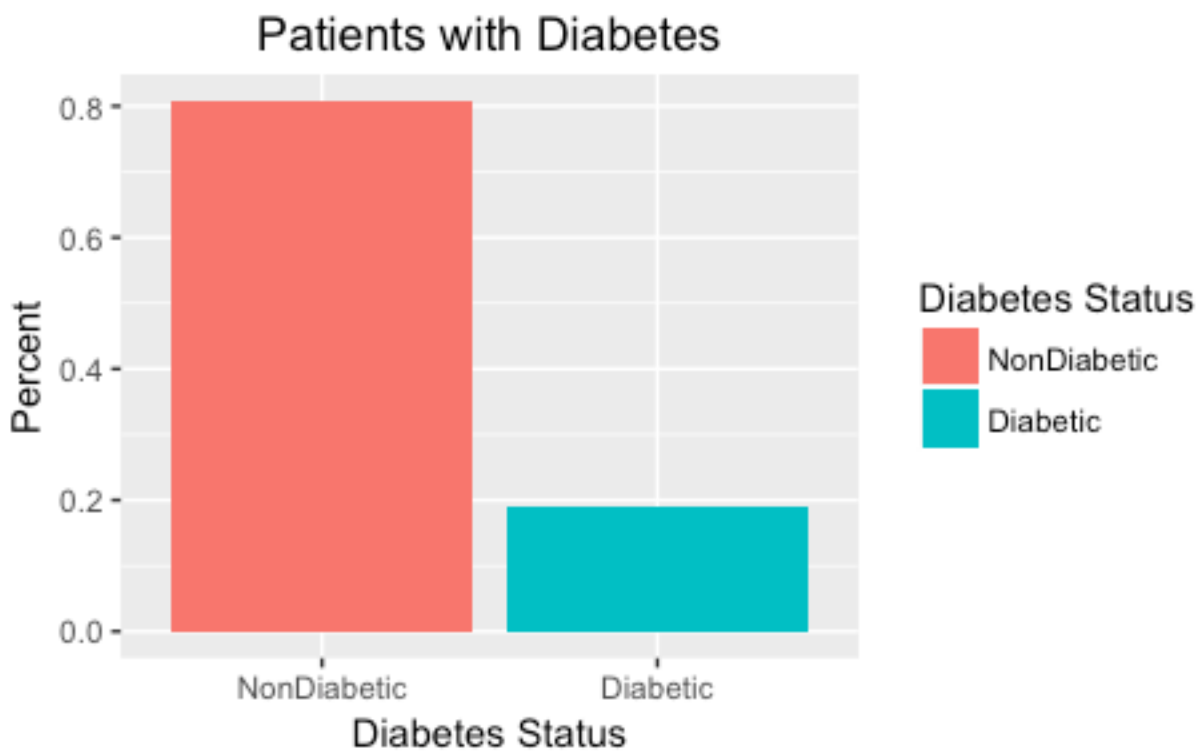


Data Cleaning

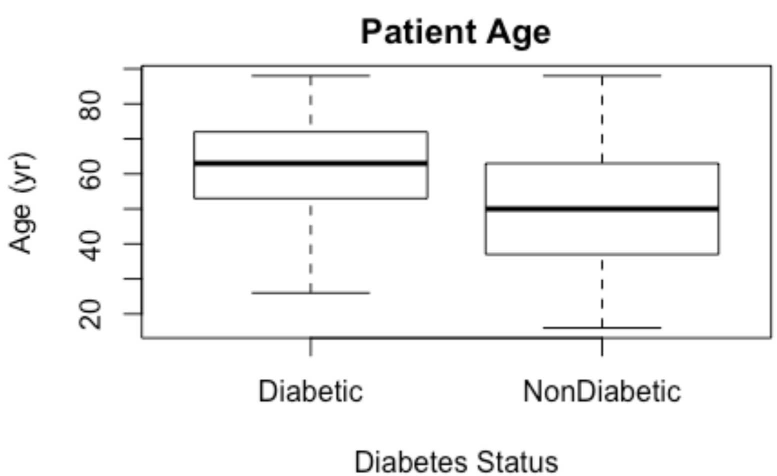
- Load data from CSV files
- Join data from related tables
- Cluster diagnosis data into diagnosis categories
- Map medication NDC codes to medication names
- Filter data
  - Remove errors
  - Isolate relevant data
- Transform data types
- Derive data
  (e.g. pulse pressure = systolic – diastolic)

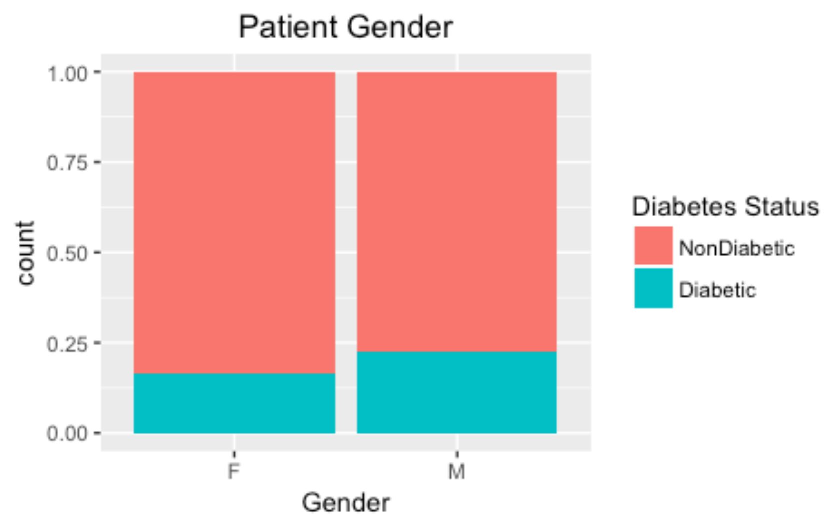
Exploratory Analysis

Baseline Ratio of Diabetic Patients

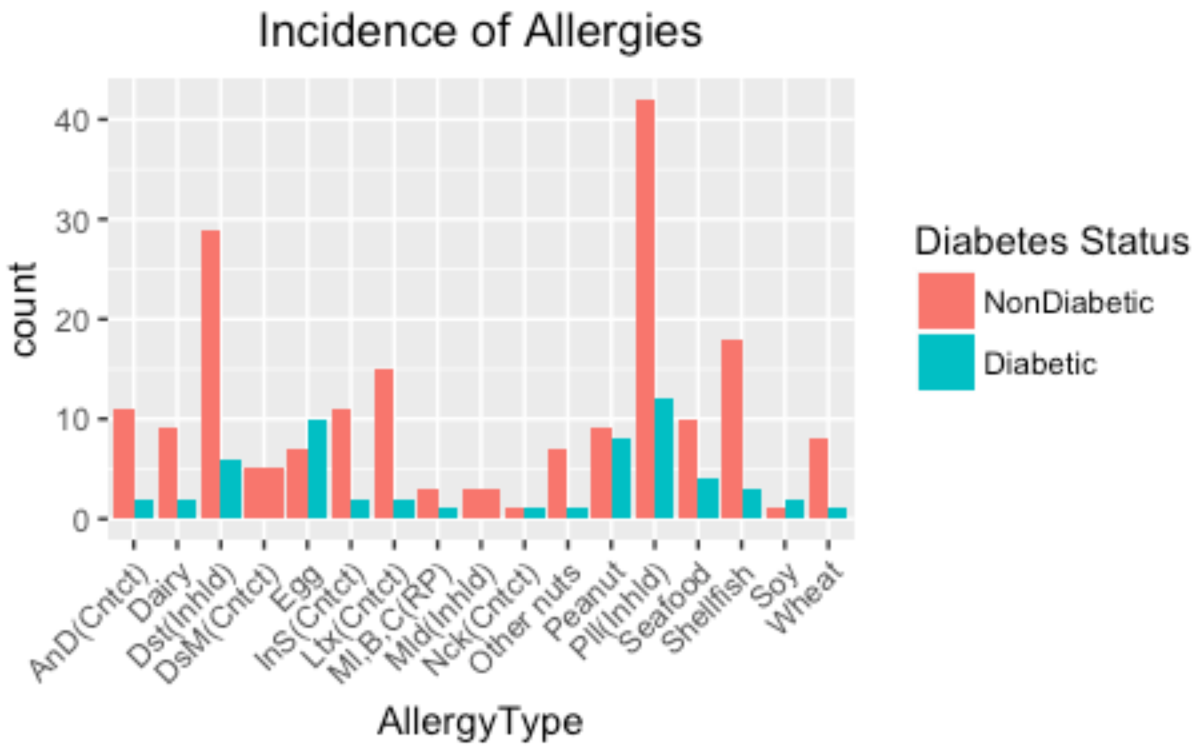


Age & Gender


Patient Age
Age (yr)
20
40
60
80
Diabetic
NonDiabetic
Diabetes Status

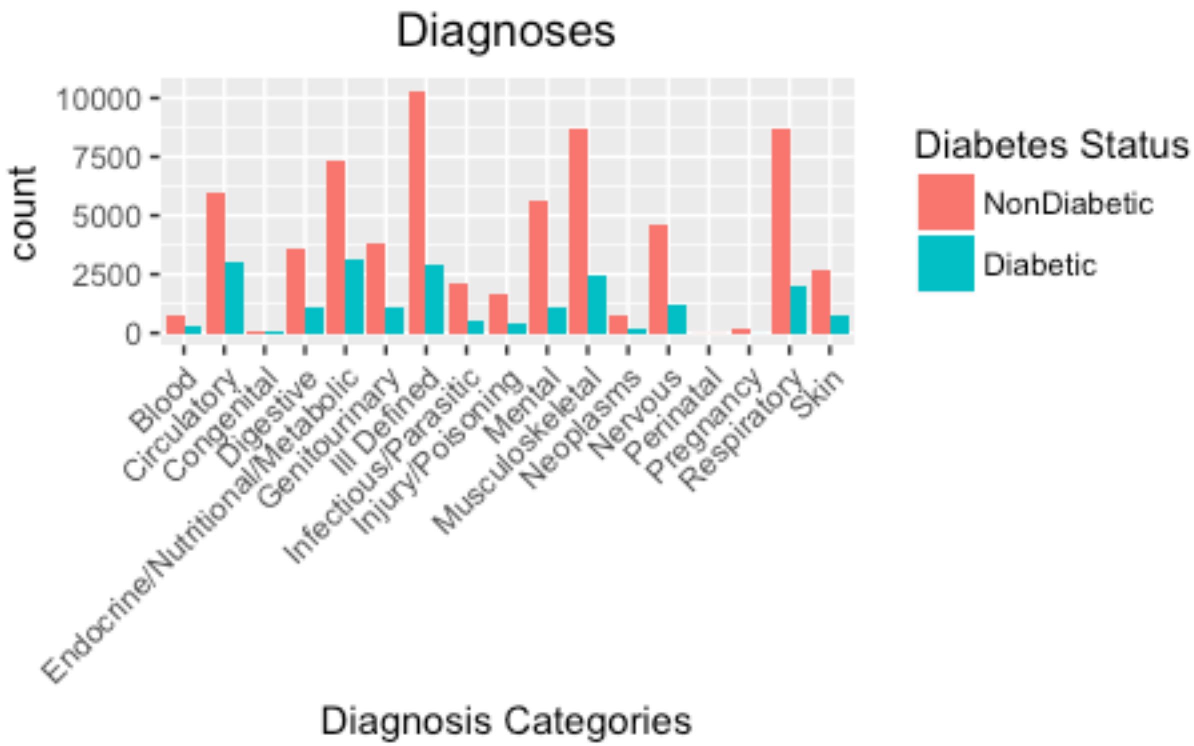

Patient Gender
count
0.00
0.25
0.50
0.75
1.00
F
M
Gender
Diabetes Status
NonDiabetic
Diabetic

Allergies



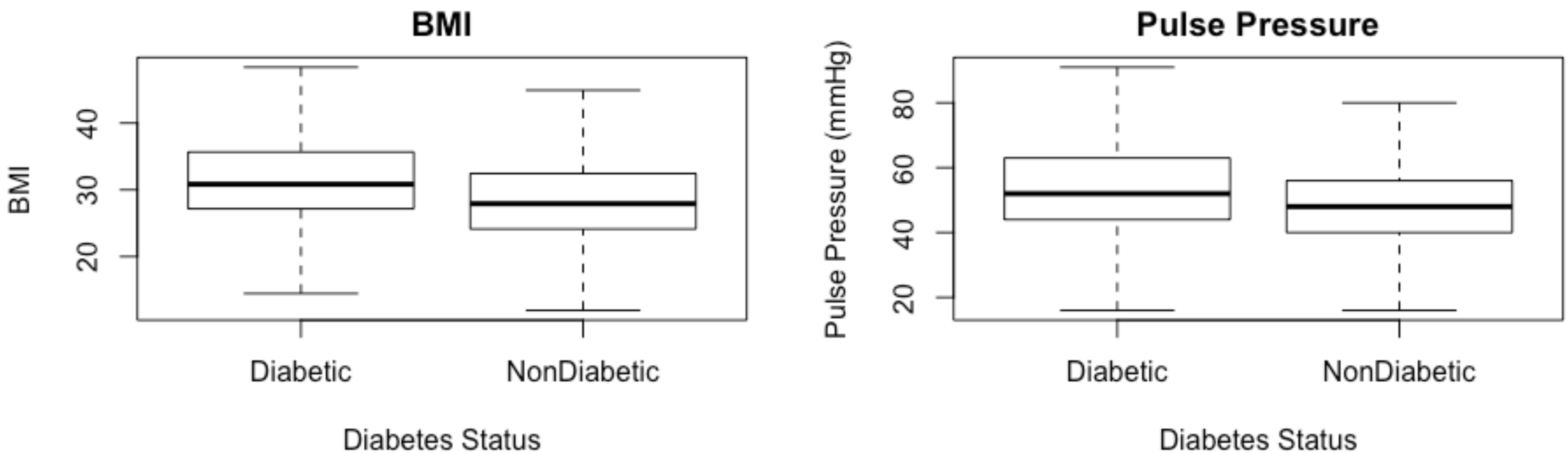
Incidence of Allergies
count
40
30
20
10
0
AnD(Cntct)
Dairy
Dst(Inhld)
DsM(Cntct)
Egg
InS(Cntct)
Ltx(Cntct)
Ml,B,C(RP)
Mld(Inhld)
Nck(Cntct)
Other nuts
Peanut
Pll(Inhld)
Seafood
Shellfish
Soy
Wheat
AllergyType
Diabetes Status
NonDiabetic
Diabetic

Diagnoses

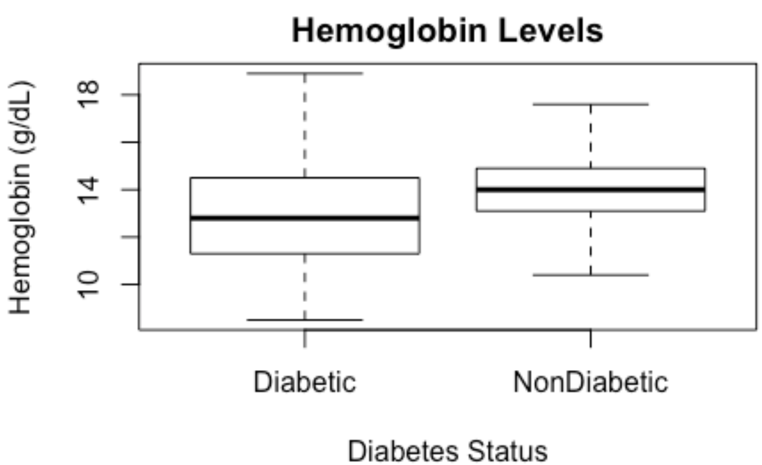


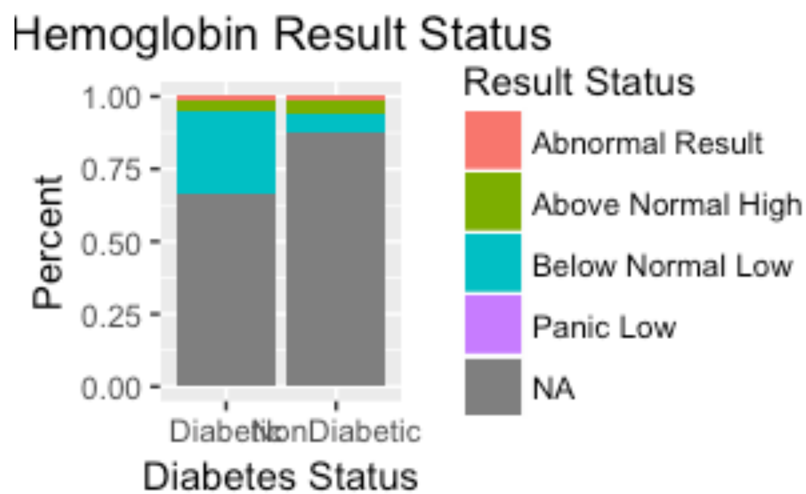
Diagnoses
count
10000
7500
5000
2500
0
Blood
Circulatory
Congenital
Digestive
Endocrine/Nutritional/Metabolic
Genitourinary
Ill Defined
Infectious/Parasitic
Injury/Poisoning
Mental
Musculoskeletal
Neoplasms
Nervous
Perinatal
Pregnancy
Respiratory
Skin
Diagnosis Categories
Diabetes Status
NonDiabetic
Diabetic

Transcript Data: BMI & Pulse Pressure

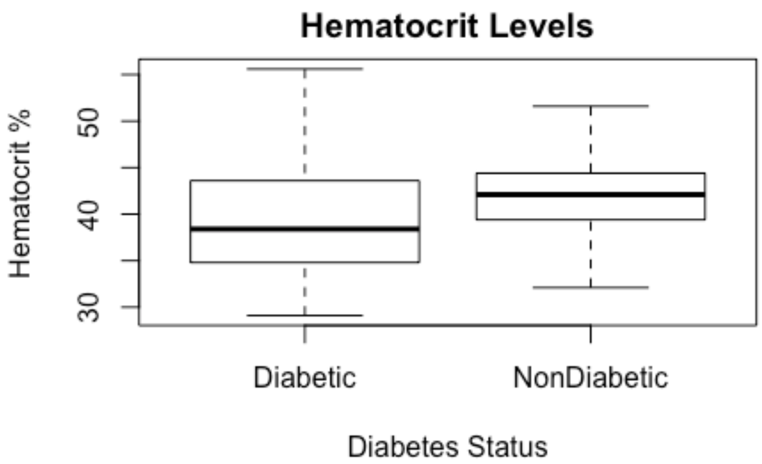


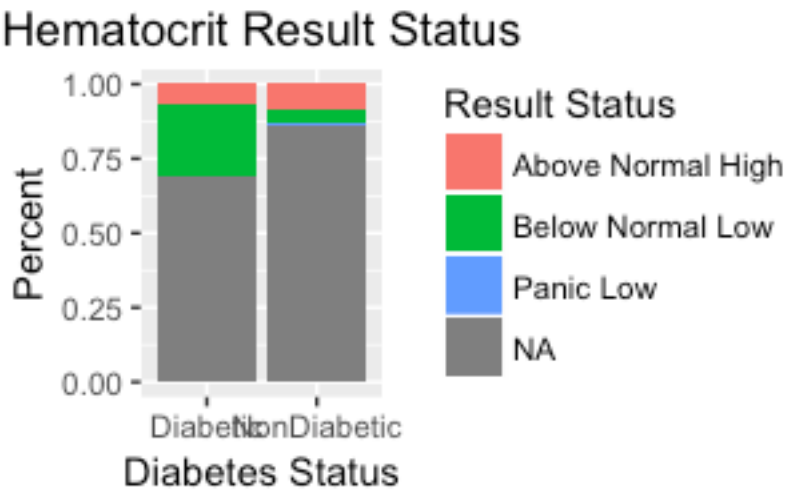
Labs: Hemoglobin Levels



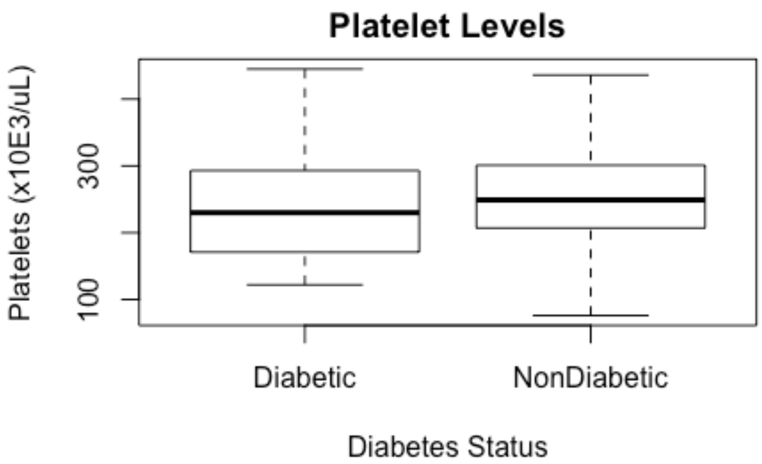


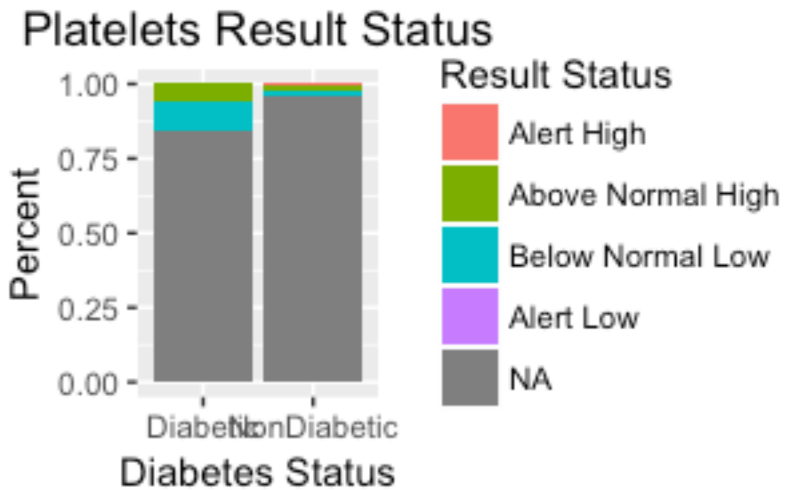
Labs: Hematocrit Levels





Labs: Platelet Levels





Predictive Models

Promising Variables

- Age
- Gender
- Allergies
- Diagnosis Categories
- BMI
- Pulse Pressure
- Hemoglobin Levels
- Hematocrit Levels
- Platelet Levels

Considerations

- Mix of continuous and categorical data
- Multiple observations per patient
- Disconnected data

Random Forests Predictive Model

- Can be used with mixed data types
- Can be used with non-linear data relationships
- Evaluation
  - Training/Test Predictions: Confusion Matrix
  - 10-Fold Cross-Validation

Lab Results Data Predictive Model

- Isolate Hemoglobin, Hematocrit, Platelets data
- Imputation to estimate missing values
- Split into training and test subsets
- Random Forest Model:

dmIndicator ~ age + gender + hemoglobin + hematocrit + platelets

Lab Results Data Predictive Model

- Confusion Matrix: 88% Accuracy

		Predicted	
Actual		0	1
	0	216	3
	1	27	3

- Cross Validation:
  - Accuracy: 89%, Kappa: 31%

Diagnosis Data Predictive Model

- Separate diagnosis category factors into columns
- Aggregate data into single row containing all diagnoses
- Include transcript data
- Split into training and test subsets
- Random Forest Model:

dmIndicator ~ age + gender + BMI +
pulse pressure + endocrine +
circulatory

Diagnosis Data Predictive Model

- Confusion Matrix: 79% Accuracy

	Predicted		
Actual		0	1
	0	9,637	317
	1	2,470	616

- Cross Validation:
  - Accuracy: 81%, Kappa: 43%

Combined Predictive Model

- Combine allergy, diagnosis & transcript Data
- Too many medications for model
- Split into training and test subsets
- Diagnosis data not significant in this model
- Random Forest Model:

dmIndicator ~ age + gender + allergies

Combined Predictive Model

- Confusion Matrix: 95% Accuracy

	Predicted		
Actual		0	1
	0	782	1
	1	39	25

- Cross Validation:
  - Accuracy: 99%, Kappa: 95%
  - Overfit?

Discussion & Conclusion

Limitations

- Lab data not linked to doctor visit transcript data
- Medication data too granular (need to chunk medication into categories)
- Possible overfit of combined model
  - Too many factor levels?

Conclusions

- Contributors to predictive models
    - Age & gender
    - BMI & pulse pressure
    - Hemoglobin, hematocrit, platelets
    - Endocrine & circulatory diagnoses
    - Allergies
- Recommendations
    - Further study
    - Collect more data around these factors
    - Allergies may be a novel area of research