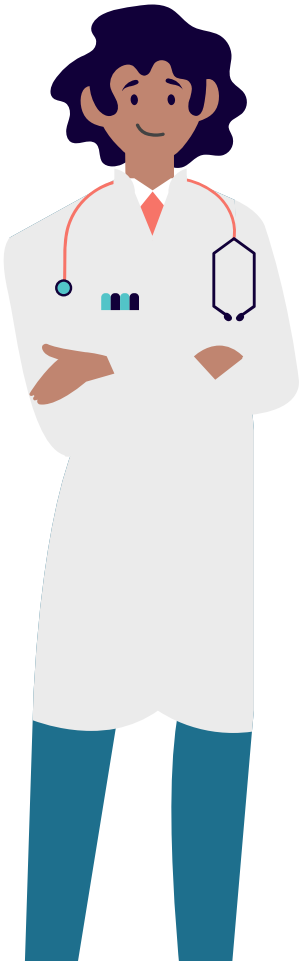


Predicting Diagnoses from Drug-Related Text Data

MSiA 414 - Fall 2021
Brian Lewis

How Medical Coding Works

1

Diagnosis

A patient is admitted and given a diagnosis after assessing symptoms

2

Drug Regimen

The patient is given one or more prescription drugs to treat the diagnosis

3

Insurance Billing

The doctor submits billing codes for diagnosis using ICD-9 / ICD-10 standards to the patient's insurance company

But what if you switch doctors?

Doctor:

"We don't have your old medical records. What medications do you take?"

Doctor:

"Hmm, I haven't heard of albuterol before. Let me check our NLP system!"

Doctor:

"It looks like your last doctor probably diagnosed you with chronic bronchitis."

Patient:

"I take albuterol, which I think is supposed to help my breathing."

Patient:

"That would be helpful, I don't remember my prior diagnosis."

Dataset: MIMIC-IV

40,000+

Patients

All in-patient, hospital admissions from the Beth Israel Deaconess Medical Center in Boston, MA

6,000+

Drugs

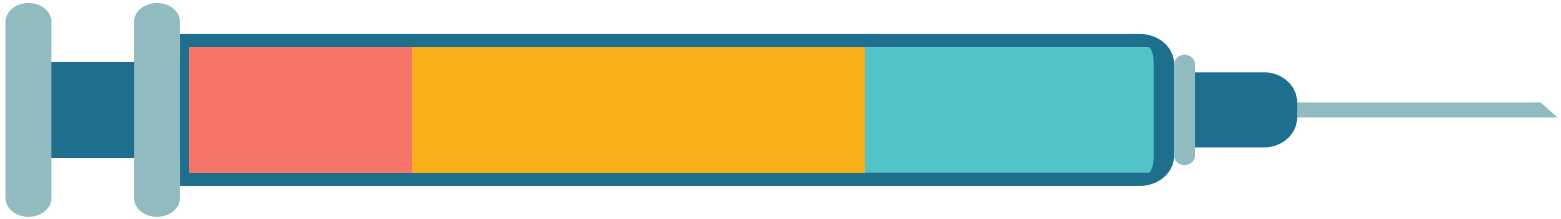
The data include a variety of prescription drugs used in conjunction with patient stays in the hospital

27,000+

Diagnoses

The data include tens of thousands of distinct diagnoses. Most patients receive over 10 diagnoses.

Data Processing



1

Diagnosis per patient

Filtered the data to one diagnosis per patient

50

Most frequent diagnoses

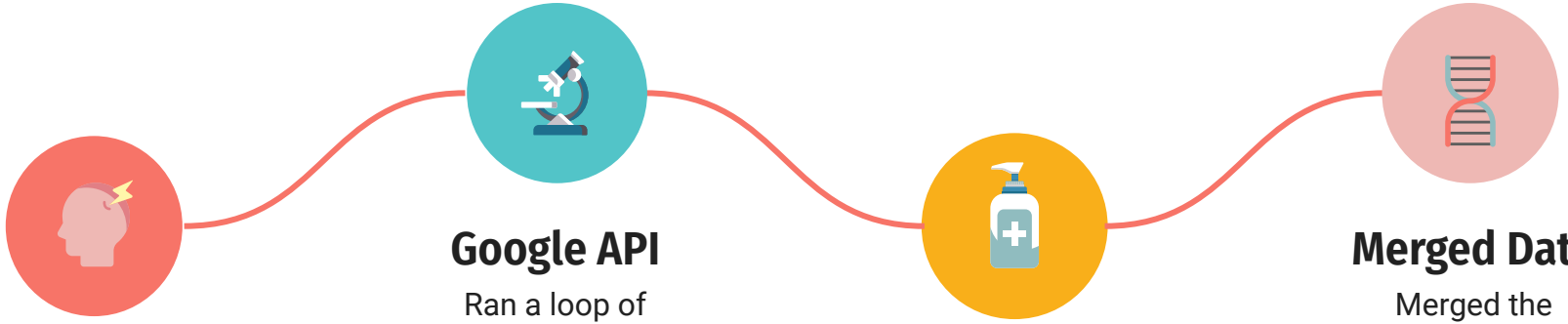
Filtered again to top 50 most frequent diagnoses

2.21 M

Total Observations

After cleaning and processing, result was 2.21 million observations

Where did the text data come from?



Unique Drugs

Filtered to unique list of drug names

Google API

Ran a loop of "***What is <drug name> used for***" into Google's "Custom Search" API

Cleaned JSON

Cleaned and parsed the text portions of the Google API responses

Merged Data

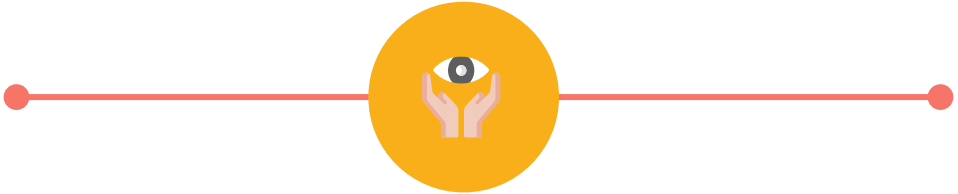
Merged the cleaned JSON text data about each drug back on to the associated diagnosis

NLP Process

Tokenize

All text from Google API parsed JSON was cleaned and tokenized

Balance



Minority classes were balanced with random oversampling

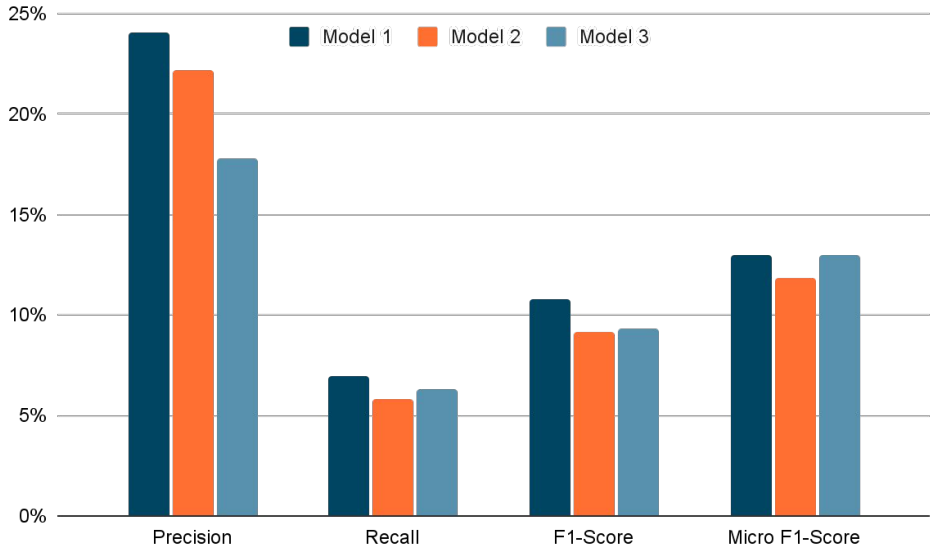
Logistic Regression

Logistic regression classified each drug's text with a particular diagnosis and probability

App Demo!

Link to Flask App

Model Performance



Low Performance

No other academic work has tackled this precise problem, but the percentages are low.

Complex Problem

Messy text data, imbalanced classes, and narrow focus on inpatient admissions and diagnoses make this problem very complicated.

Executive Summary

✓	The MIMIC-IV dataset contains useful linkage data between prescriptions and diagnoses
✗	Most of these data are related to pregnancy admissions and only focus on hospital visits
✓	Google's Custom Search API allowed for gathering unstructured text data related to each drug
✓	Logistic Regression was able to create a somewhat meaningful classifier for 50 outcome classes
✗	Model performance was low; cleaner data and better methods would be useful in future work