Using Patient Medical Journey Data to Identify Undiagnosed Heart Disease

Ashok Kumar Reddy, Caitlin Hassey, Kipp G Spanbauer, Pankaj Shrestha, Sean A Lee, Stephanie Hsieh

Motivation/Introduction

Heart Disease remains a significant cause of death in the United States. Chronic Heart Disease accounted for approximately 12.6% of deaths in the United States causing 360,900 deaths in 2018. Furthermore, data for the years 2005 to 2014 show the estimated annual incidence of heart attack in the United States were 605,000 new attacks and 200,000 recurrent attacks. Average age at the first occurrence of heart attack was 65.6 years in males and 72.0 years in females. The estimated direct and indirect cost of heart disease from 2017 to 2018 (average annual) was $228.7 billion in the United States [1].

Using Medical Claims data, we created machine learning models that the healthcare systems and/or researchers could use to predict the risk of having a heart disease based on the historical claims or medical visits of people and thus identify patients during the early stages of heart disease. Visualizing the models and data in the context of geography and other demographics of patients provides valuable insight and guidance in improving the ability to diagnose a patient at an early stage in addition to helping doctors build better patient care leading to better outcomes while reducing overall cost.

Approach

Our approach balances the use of powerful analytics, complexity of models, and interpretability of results. Our critical innovations include the following:

1. Novel approach to data curation, using patient medical journey data
2. An interactive U.S. map, using Tableau, to help medical providers and researchers understand where cardiovascular disease (CVD) is underdiagnosed, and the demographics of these patients

This is an expansion to provide data utility like the well-known Framingham Heart Study[2] which focuses on a small geographical area and population

We wanted to identify patients who are at high risk of developing cardiovascular disease but had not yet been diagnosed. For the context of this project, we have termed these high-risk patients as "undiagnosed".

Data Curation

- Based on our literature survey, identified key medical conditions attributing to heart disease
- Using SNOMED Controlled Terminology browser, identified concept codes[3] related to heart disease and categorized them. This was achieved by mapping concept codes to identified medical conditions and prescription history to generate a curated dataset from raw data
- The independent variables are patient demographics (location, gender, age, race, ethnicity), medical conditions (diabetic, chronic diseases, high blood pressure, obesity) and prescription history (use of Statin, use of Aspirin)
- The dependent variable is the probability of patient having heart disease
- Raw data wrangling first performed on 100K patient dataset in Postgres database and second on 2.3 million dataset leveraging PySpark on a Databricks Spark cluster

Analytics Models Design

- 100K patients were randomly selected from 2.3 million patient dataset for a training set
- With this, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, K Nearest Neighbor, Gaussian Naïve Bayes models were fitted in R and Python (Scikit-learn)
- Predictions of patient's cardiovascular disease status was created on the remaining curated 2.2 million patients
- For Logistic Regression, LASSO and Stepwise Regression variable selection techniques were experimented
- The most significant independent variables were diabetes, chronic diseases, high blood pressure, obesity, age
- The best model had the accuracy of around 74%

Experiments and Results

Data Curation

- We attempted to select statistically significant factors in the data by using a one-hot encoding approach of concept codes. This resulted in a sparse matrix which would not yield the best results for our models. Ultimately, we chose multiple concept codes based upon the literature survey to build independent predictors of heart disease

Models

- Explored Logistic Regression, Random Forests, Decision Trees, Support Vector Machine, Gaussian Naïve Bayes, and K Nearest Neighbor models using an 80/20 training/testing split on 100K patient records
- Performed data exploration, checked for multicollinearity, and assessed model goodness of fit
- Predicted outcomes using remaining 2.2M records (from the total 2.3M patient data set) and evaluated model quality using confusion matrices and looking at the accuracy, sensitivity, and specificity
- Our literature review supports that the cost of over diagnosing cardiovascular disease (false positive) is less costly than underdiagnosing (false negative), and thus we experimented with including a notion of cost into our analysis – estimating that underdiagnosing is 2x costly as over diagnosing
  - To support this, for our Logistic Regression model, explored 40%, 50% and 60% classification thresholds, and ultimately landed on a 40% threshold that would be more conservative in diagnosing cardiovascular disease and reducing the overall cost of misclassification in the model

	Accuracy	Sensitivity	Specificity	Cost (Even)	Cost (2x)
Logistic Regression 40% Threshold	0.7396	0.8109	0.6684	579,889	790, 158
Random Forest	0.7386	0.7769	0.6900	582,138	859,908
Decision Trees	0.7380	0.7743	0.6912	583,440	866,638

Table: Model accuracy and cost for top 3 options

- Performed dimensionality reduction with Principal Component Analysis (PCA) to two principal components to plot the decision boundaries and visualize how the model predicts cardiovascular disease

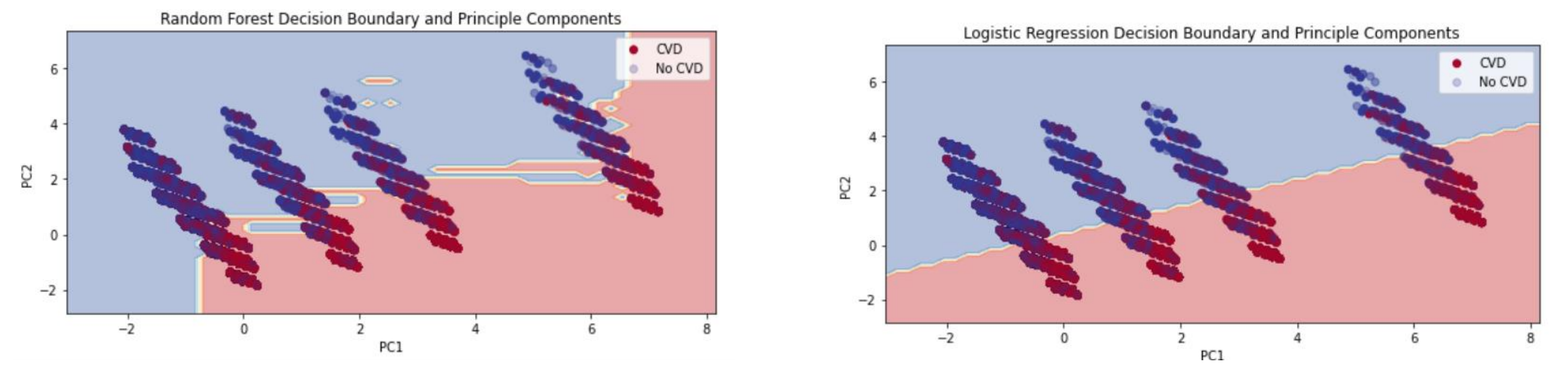


Figure: Random Forest and Logistic Regression decision boundaries to visualize model accuracy by comparing predictions to actual patient cardiovascular disease (CVD) status. Red zone: model's prediction for CVD, Blue zone: prediction for no CVD, Red datapoints: patients with CVD, Blue datapoints: patients with no CVD

Tableau

- Experimented with two different dashboards to assess questions researchers might want to answer:
  - If a researcher wants to know what overall percentage of a selected age group based on the total state population, they can use the "Undiagnosed Overview by Total Population" dashboard
  - If they want to know the percentage undiagnosed of a particular selected cohort, the researcher could use the "Undiagnosed Overview by Selected Population & Factors" dashboard

Our approach to identification of undiagnosed heart disease differs in following ways compared to similar initiatives

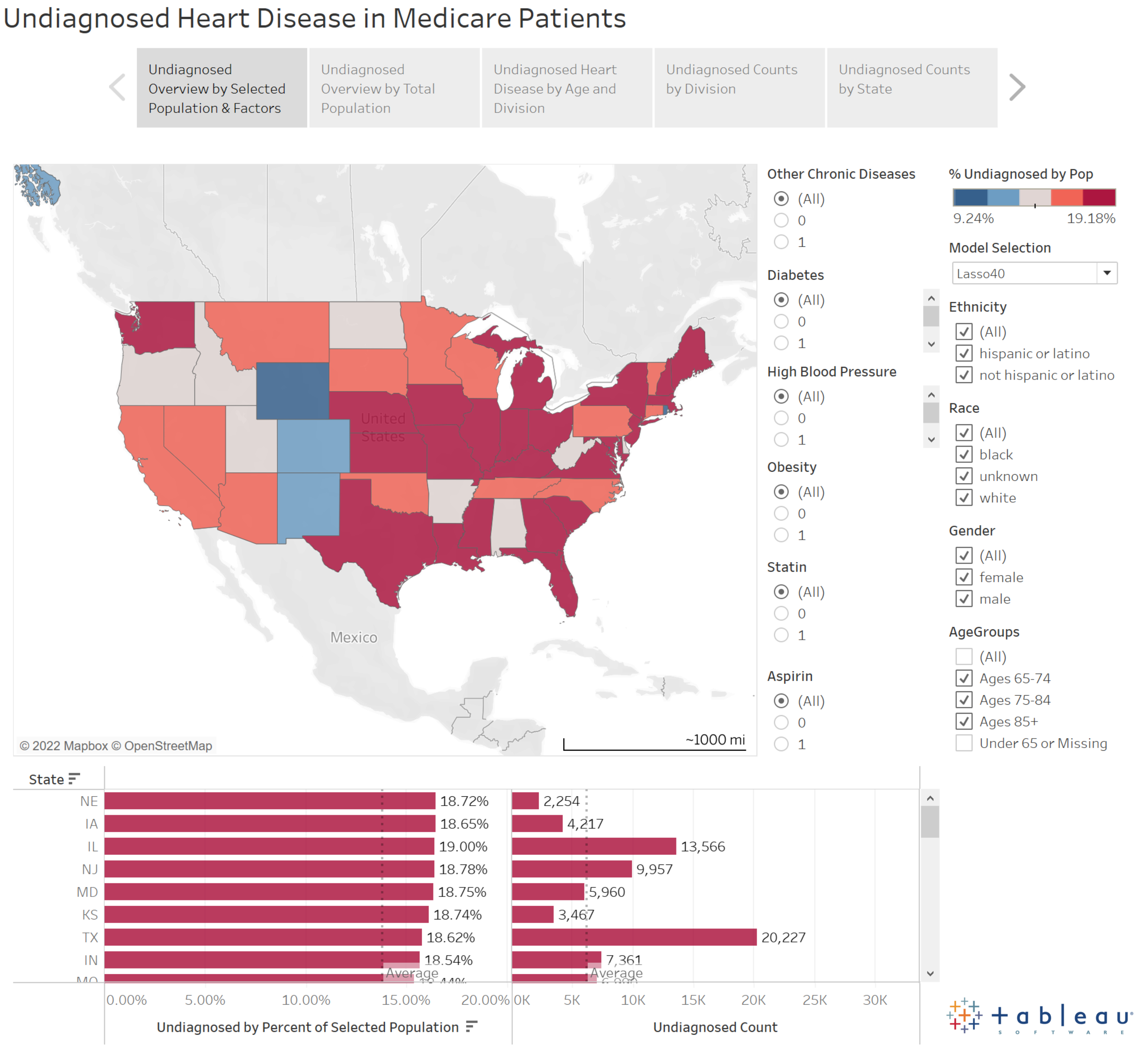
- Use of Medicare claims data to build and train multiple models to predict patient's chances of heart disease is different than previous studies like the Framingham Heart Study. Framingham Study primarily used Logistic Regression on a much smaller sample size, our prediction methods utilize a more comprehensive regression analysis and multiple classification models. Our models pull prior medical data from a rolling database and do not require in-person consultation of participants to acquire the data. This approach is much quicker and less expensive means to pull a dataset that is of magnitudes larger than prior studies
- Novel approach to data curation using anonymized Medicare claims data. Using Observational Medical Outcomes Partnership (OMOP) patient medical journey data model, we then used SNOMED to identify heart disease concept codes and their hierarchical relationship
- Experimented with including a notion of cost into our analysis, estimating that underdiagnosing of cardiovascular disease is 2x costly as over diagnosing
- For Linear Regression model, LASSO and Stepwise Regression was performed for independent variable selection to streamline the important variables, which helps with better interpretability of prediction and wider adoption

References:
1. Heart Disease and Stroke Statistics - 2022 Update from American Heart Association
2. Framingham Heart Study
3. SNOMED Browser and Concept Codes
4. 2.3 million patients dataset
5. 100 thousand sample patients dataset

Data

Data Source	Our dataset was from Medicare Claims SynPUF data for 2008-2010 in the Observational Medical Outcomes Partnership (OMOP) Common Data Model from Observational Health Data Sciences and Informatics (OHDSI) with 2.3 million patients aged 65 and older
Data Characteristics	~ 480 million rows (including all the supporting datasets) ~ 18 GB on disk space
Download Links	2.3 million dataset[4] 100 thousand sample dataset[5]

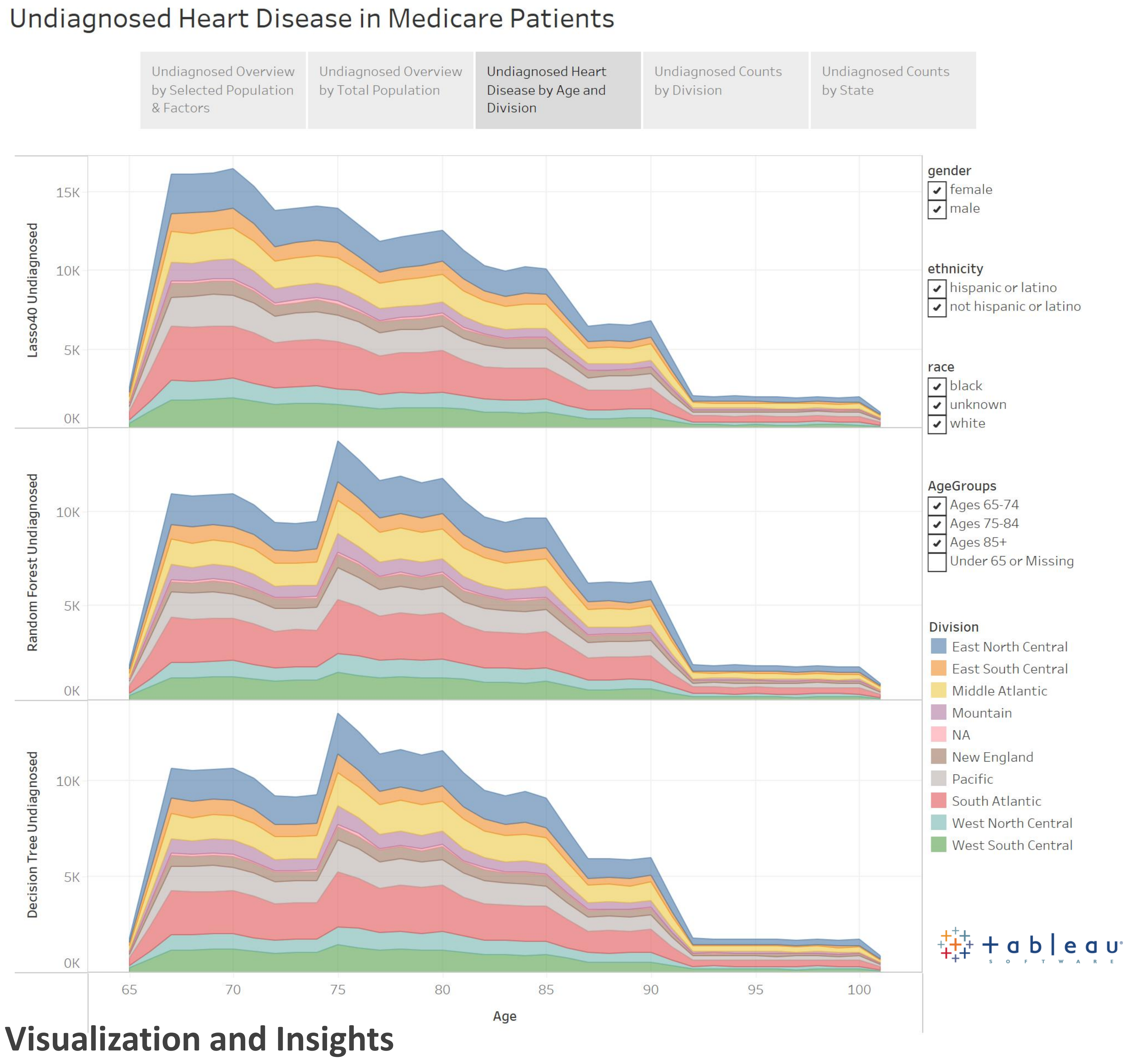
Table: Data source details


Undiagnosed Heart Disease in Medicare Patients

Undiagnosed Overview by Selected Population & Factors
Undiagnosed Overview by Total Population
Undiagnosed Heart Disease by Age and Division
Undiagnosed Counts by Division
Undiagnosed Counts by State
United States
Mexico
© 2022 Mapbox © OpenStreetMap
~1000 mi
% Undiagnosed by Pop
9.24%
19.18%
State
Undiagnosed by Percent of Selected Population
Undiagnosed Count
Average
0.00%
5.00%
10.00%
15.00%
20.00%
0K
5K
10K
15K
20K
25K
30K

Other Chronic Diseases: ● (All) ○ 0 ○ 1
Diabetes: ● (All) ○ 0 ○ 1
High Blood Pressure: ● (All) ○ 0 ○ 1
Obesity: ● (All) ○ 0 ○ 1
Statin: ● (All) ○ 0 ○ 1
Aspirin: ● (All) ○ 0 ○ 1
Model Selection: Lasso40
Ethnicity: ☑ (All) ☑ hispanic or latino ☑ not hispanic or latino
Race: ☑ (All) ☑ black ☑ unknown ☑ white
Gender: ☑ (All) ☑ female ☑ male
AgeGroups: □ (All) ☑ Ages 65-74 ☑ Ages 75-84 ☑ Ages 85+ □ Under 65 or Missing

State	Undiagnosed by Percent of Selected Population	Undiagnosed Count
NE	18.72%	2,254
IA	18.65%	4,217
IL	19.00%	13,566
NJ	18.78%	9,957
MD	18.75%	5,960
KS	18.74%	3,467
TX	18.62%	20,227
IN	18.54%	7,361


Undiagnosed Heart Disease in Medicare Patients

Undiagnosed Overview by Selected Population & Factors
Undiagnosed Overview by Total Population
Undiagnosed Heart Disease by Age and Division
Undiagnosed Counts by Division
Undiagnosed Counts by State
Lasso40 Undiagnosed
Random Forest Undiagnosed
Decision Tree Undiagnosed
0K
5K
10K
15K
Age
65
70
75
80
85
90
95
100
Division
East North Central
East South Central
Middle Atlantic
Mountain
NA
New England
Pacific
South Atlantic
West North Central
West South Central

gender: ☑ female ☑ male
ethnicity: ☑ hispanic or latino ☑ not hispanic or latino
race: ☑ black ☑ unknown ☑ white
AgeGroups: ☑ Ages 65-74 ☑ Ages 75-84 ☑ Ages 85+ □ Under 65 or Missing

Visualization and Insights

Multiple dashboards were created in Tableau and combined using Tableau's story feature

"Undiagnosed Overview by Selected Population & Factors" dashboard filters the Medicare population that our models have identified as being at risk of heart disease but are yet to diagnosed. This analytics can be performed at geographic state level with ability to filter data by different independent variables, all the prediction models mentioned, and their various cohorts. The bar charts at the bottom ranks US States by undiagnosed percent of selected population or the total count. A researcher can quickly and easily identify the US States with the highest potential of undiagnosed heart disease population based on demographics and medical criteria. This can lead the research team to formulate a plan of action in getting those at risk prioritized health screening for diagnosing heart disease.

Similarly, "Undiagnosed Heart Disease by Age and Division" shows undiagnosed heart disease population by age and prediction models across different geographic divisions with the ability to filter by cohorts.

There are additional dashboards that provide different views of this data based on questions researchers want to ask and building new ones that don't already exist should be a trivial effort assuming the data is present