# Using Patient Medical Journey Data to Identify Undiagnosed Heart Disease

## Introduction

Heart disease kills more people in the United States annually than any other disease and reduces quality of life for patients. Identifying heart disease patients can be complicated, and with increasing demands on medical providers, there may be missed opportunities diagnosing patients. If diagnosed early enough, managing heart disease through lifestyle changes and better medications can lead to a longer, healthier life for patients.

## Problem Definition

Using healthcare claims data, we created machine learning models that the healthcare system and/or researchers could implement to help predict patients who are at risk for heart disease and identify undiagnosed patients with heart disease at earlier stages. Visualizing this in the context of geography and other demographics provides an important framework, and thus this is a key piece of our analysis. Improving patient diagnosis and medication treatment will help doctors build better patient care leading to better outcomes while reducing overall cost.

## Review of Literature

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 (Tsao et al, 2022).

While primary prevention is the first course of action in any health issue, medical professionals often believe that taking secondary prevention is just as important. (Müller-Riemenschneider et al, 2010) states that there is evidence that secondary prevention, particularly without the use of drugs as treatment (nonpharmacological) can be an effective intervention in cardiovascular disease (CVD). Improving upon secondary prevention measures will allow for better treatment of heart disease patients since conditions are not yet severe; however, identifying at-risk patients is not straight-forward.

Homing in on risk factors will help determine what course of action to take in secondary prevention. A three-year study (Khot et al, 2003) focused specifically on four risk factors' impact on coronary heart disease (CHD) rates: smoking, diabetes, hyperlipidemia and hypertension. (Lida et al, 2003) identified that blood pressure is highly correlated with mortalities from stroke and CVD. (Alm-Roijer et al, 2004) investigated lifestyle risk factors for heart disease and if any changes will help medication treatment goals and adherence to drug therapy. Finally, from 1980-2008, 37 clinical studies from 40 countries were pulled to show the underlying risk factors for various heart diseases and conditions across six regions of the world (Khatibzadeh et al, 2013). This gave a better understanding of how different factors of CVD affected different regions and what course of action to take in prevention.

Heart disease is well studied, and with the emergence of AI, analytical techniques have become more prevalent. (Karunathilake et al, 2018) analyzes secondary prevention for fighting CVD and discusses ways in which technology can aid in early diagnosis. (Papez et al, 2021) analyzes clinical data for identifying heart failure visualized through mapping densities of risk factors like geography and age groups. (Alaa et al, 2019) created a new machine learning model called AutoPrognosis to predict CVD risk based on 473 variables and compared the accuracy of its predictions to the seven conventional risk factors of CVD assessed in the Framingham score and Cox proportional hazards models. (D'Agostino et al, 2008) developed sex-specific multivariable risk prediction algorithms that can be used to help quantify risk so that it can guide preventive care and used by primary care physicians to assess patients and their risk of developing CVD. Currently, there are multiple methods of unsupervised and supervised machine learning used in cardiovascular medicine (Johnson et al, 2018). Deep learning and neural networks are difficult to interpret, with (Bai et al, 2018) taking a new deep learning approach to predict the next patient visit and main diagnosis. There is a tradeoff between accuracy and interpretability of the deep learning model of which one should be mindful.

The Framingham Study is the longest leading study in heart disease and epidemiological studies. (Wilson et al, 1998) used linear and logistic regression to investigate the relationship between different risk factors, especially blood pressure and cholesterol categories, on CVD risk. This study was very influential on our approach, though it is limited to a specific geography. In addition, we considered many other machine learning models, many of which overlap with the discussion by (Parthiban & Srivatsa, 2012), where patient attributes and medical history can be used to classify individuals at risk for heart disease. Similar to (Choi et al, 2020), we can confirm the accuracy of our model by measuring the concordance rate between our machine learning diagnosis and diagnosed patients with CVD.

Many health studies can be challenging to interpret as they may have patients self-report results, increasing the risk of inconsistent data. From 1982-2003, a study (Glynn et al, 2005) correlated risk factors to the first occurrence of their CHD, stroke, or venous thromboembolism. The results were conclusive in detecting change in CHD, but the data was primarily self-reported data and only consisted of Caucasian male participants from the U.S. Having an unbiased view of the data will be an important foundation of prediction.

Finally, it was critical that we consider CVD from a wider societal lens. (Fuller et al, 2019) looks at how different socioeconomic inequalities affect health conditions and found that people living in lesser conditions with less access to healthcare were more prone to disease. This highlights the importance of geography in this research. Additionally, governments are a large stakeholder in healthcare and (Benjamin et al, 2019), on behalf of the American Heart Association, estimate that the annual cost of CVD is $350b (direct and indirect costs) and are projected to more than double from 2015 to 2035. (Ariza et al, 2010) argues that specifically for patients with diabetes, complications like CVD make up a significant proportion of health care spending.

## Proposed Methods and Intuition

Our methods balance the use of powerful analytics, complexity of models, and interpretability of results. Our critical intuitions include:

1. Using patient medical journey data as a novel approach to data curation.
2. Creating an interactive U.S. map, using Tableau, to critically help medical providers and researchers understand where CVD is underdiagnosed, and the demographics of these patients.
    a. This is a key expansion of the well-known Framingham study, 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

Our approach is better than the state of the art because of our novel approach of how we curated our data. To accomplish this, we curated a set of anonymized Medicare claims data using SQL: raw data wrangling first performed on 100K patient dataset in Postgres database and secondly on a 2.3 million patient dataset leveraging PySpark on a Databricks Spark cluster. From there, we built models in R and Python and displayed our results using a customized, interactive visual display in Tableau.

The data asset we chose for our project is the 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. Each table has a coded value for representing the information which can be decoded using the concept tables. Another feature of the data is that diagnosed disease states (the condition_era table), symptoms collected during physician visits (the observation_era table), and medications which have been prescribed to patients (the drug_era table) use a hierarchical, controlled terminology called Systemized Nomenclature of Medicine (SNOMED).

SNOMED has codes mapped for all the data in these tables and is designed to have a hierarchical structure. Using freely available web-based SNOMED browser, a researcher can look up codes and understand the relationships between them. For example, searching for “myocardial infarction” (commonly referred to as a heart attack) will show the concept code 22298006 which has 14 children codes which are more specific. Of those 14 children codes, four of them have their own children codes and families. This is the foundation for assembling Patient Journey data.

### Geographical Visualization

The Framingham study data is a smaller subset of a specific geographical location that primarily used regression analysis. While insightful, our second key innovation expands our population and statistics nationally and presents an easy to understand and compelling method. We used our machine learning model to identify geographical areas where there may be gaps in medical care for patients who may have undiagnosed heart disease. Our team would like to take similar ideas from the literature findings one step further by presenting our results using Tableau as a visualization platform. This will also be state of the art because it will allow researchers to work interactively toward solutions to identify potential patients who may have undiagnosed heart disease based upon models built from heart disease patients. This innovation could drive the use of medical journeys in future studies and lead to impactful outcomes. We hope that this map could help raise awareness of heart health and cardiovascular disease.

## Analytical Methods

As discussed, curating the data is the first major step. From there, we leveraged R and Python to run experiments and perform analysis.

- We built our training set by randomly selecting 100K patients from our 2.3 million patient dataset and predictions of patient's CVD status was created for the remaining 2.2 million patients.
- Using R to build logistic regression analyses to find the most important factors for predicting patients at risk of heart disease. Models are subsequently used to identify patients likely to have CVD who are currently undiagnosed.
- Using Scikit-learn in Python to analyze the performance of multiple models like Support Vector Machine, Naive Bayes, K nearest neighbor, Logistic Regression, Random Forest, and Decision Tree classifier models to identify those who are undiagnosed with CVD.

Visualizing the above analyses in Tableau, researchers will be able to compare models and explore cohorts to determine a recommended approach for identifying undiagnosed CVD.

## Experiments and Evaluation

Our model's goal is to pre-empt the risk of heart disease in a patient's health journey using medical claims records. The main question that our models are trying to answer is: how many patients are potentially being misdiagnosed or could they have the potential risk of developing some form of heart disease based on other health conditions? Sub-questions to help us answer our main question include: how can the patient journey data provide new insights into CVD and target patients who are at high risk of heart disease? How can our models predict a better accuracy than those that exist? How can we minimize cost of underdiagnosing patients? To help understand the flow of our experiments, we categorized our process into three sections: data curation, model analysis, and data visualization. We built many different models using our smaller dataset and trained and tested all the models to find the most accurate. We then applied our models to the larger test dataset of 2.2 million patients.

## Data

Based upon our literature survey, we searched for "heart disease" and created a heart disease response variable based upon 98 relevant concept codes for the types of heart disease supporting our hypothesis. The concept codes we chose were related to heart disease developed in adults so for example, we excluded concepts for deformities of the heart from birth.

With our response variable created, we examined various tables to determine how to best set up the data for analysis. Tables that were relevant to our project were:

- person – contains demographic data, one row per patient
- condition_era – any disease state with which a patient has been diagnosed
- drug_era – medications that a patient has been prescribed
- concept – the key value pairs containing the code/decode relationships in all the tables
- concept_ancestor – related to the concept table, lists hierarchical relationships between codes
- location – contains US States and county geography for each person.

Our initial approach to curating the data was to include all concepts and allow the model to select the most significant factors. We built a series of tables using a One Hot Encoding method where each factor was a binary value representing whether a patient had each concept or not. This resulted in over 7,800 factors with a sparse matrix; most patients only had one or two columns with a value of a one. We knew that this was an unmanageable number that would likely not result in a meaningful model.

We decided that the best way to reduce our factors was to use the concept_ancestor table to represent concepts with parent or grandparent codes and then use the One Hot Encoding approach. This approach also did not yield the desired results because a concept code could have two, three or more concept parent codes. Therefore, for every level in a hierarchy that we moved up, the reduction of factors was somewhere between 10% to 15%. We also began to wonder what all these ancestor codes would really represent in our analysis. Using this undirected approach of climbing all concept family trees one or two levels leaves each family of concept codes at various levels of specificity as not all trees have the same number of levels.

Finally, we determined that instead of allowing a model to determine the most significant factors in the data as a first step, the best approach would be a more direct one. So, we went back to our literature survey and built factors based upon common diseases and medications associated with heart disease: diabetes, chronic diseases other than heart disease, high blood pressure, obesity, statins, and aspirin. Our final dataset contains about 480 million rows which included all the supporting datasets, which take up about 18GB on disk space.

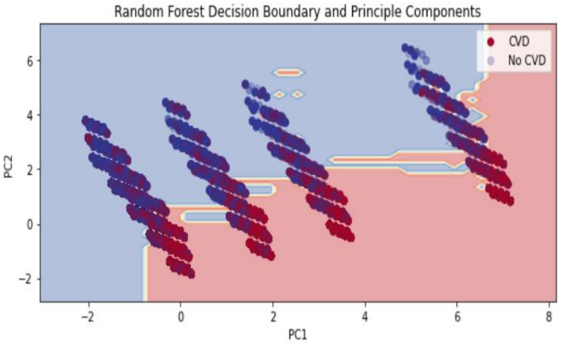
## Models

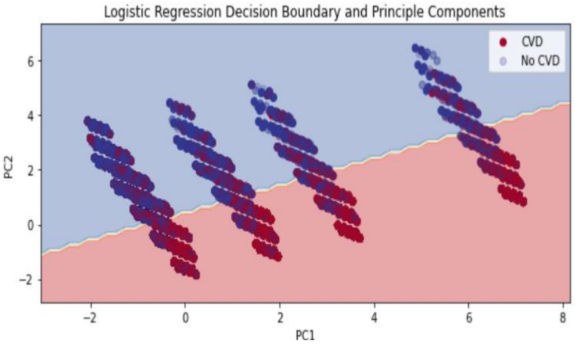
Using these factors, we were able to build a test model using logistic regression, in R, that yielded statistically significant factors from independent variables: diabetes, chronic disease, high blood pressure, obesity, statin, aspirin, age, gender, and race. We trained and tested/selected our model based on an 80/20 split on the smaller 100K dataset and used several variable selection techniques (LASSO, stepwise regression) before predicting on the remaining 2.2M records in the full dataset. We also tested the different models with various thresholds of 40%, 50%, and 60% to consider that inaccuracies in the model predictions do not have equal cost. To augment our analysis, we included data exploration, check for multicollinearity, evaluation with confusion matrices and goodness of fit analysis.

Ultimately, we chose the LASSO model at a 40% threshold. The LASSO variable selection technique produced models with the highest accuracies, and highest sensitivity as a secondary measure. We considered the cost of underdiagnosing (excess human loss) and over diagnosing (unnecessary medical expense) and felt that the cost of underdiagnosing is more significant based on research from our literature review. For this analysis, we estimated the cost of underdiagnosing to be 2x more costly, and thus we selected a 40% threshold for our model to more conservatively classify patients as undiagnosed.

| 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 |

We used similar approaches to build and analyze machine learning models in the Python Scikit-learn package: Random Forests, Decision Trees, K Nearest Neighbors, Gaussian Naïve Bayes, Support Vector Machines, and again Logistic Regression. We first took the smaller 100K dataset and split it into 80% training and 20% testing. Each model was fitted to the 80% training and then the model predicted the outputs of the 20% testing. To find the accuracies, the models' predictions were measured against the actual outputs. Using the fitted models, the CVD predictions were found on the remaining 2.2M records from the full dataset. Between all of the models, Random Forest, Decision Tree, and Logistic Regression classifiers had the highest accuracies. We believe the high accuracies of these models make sense, given the binary inputs of our independent variables. To visualize the classification models, we first performed dimensionality reduction with Principal Component Analysis (PCA) and then plotted the highest 2 principal components for each model. The decision boundaries, examples above, are a visual way of seeing how the classifier separates the predictions of "CVD" (red zone) from the "no CVD" (blue zone) and how the models' predictions compare against the actual CVD status for CVD patient (red datapoint) and non-CVD patient (blue datapoint). Logistic regression has a linear decision boundary, while random forest has non-linear and irregular decision boundary.



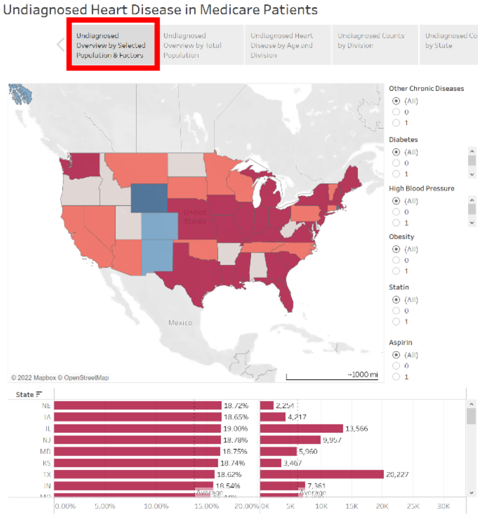


Finally, predictions across all models were compared with confusion matrices and cost analysis. Logistic regression using LASSO at a 40% classification threshold and the random forest model emerged as the optimal choices.

## Visualization

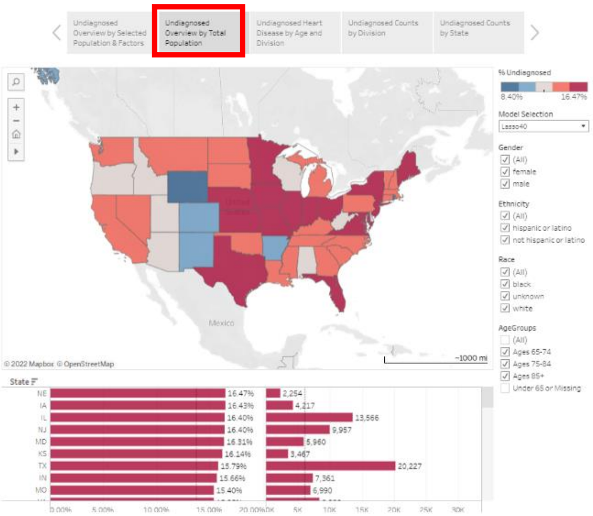
With a given set of health data from patients, our classification models can predict whether a patient is likely to have CVD and the outputs of these models can be visually represented in our Tableau map, example below. Note that in our dashboards, undiagnosed patients in our visual refers to those patients that do not have CVD according to the data, but our model predicts that they do have CVD.

We can filter our map by model, sex, race, ethnicity, age, and other factors. If hovering over a state on our map, tooltips can provide more detailed statistics such as the accuracy of each model. There are linked horizontal bar charts below the US map which can be sorted to show the highest percentage or highest count of undiagnosed participants per state. While there are many different ways our data can be visualized, we developed five different screens that best represent our data which we explore in detail below. We hope that this visualization could potentially help identify the largest need for undiagnosed heart disease populations.


Undiagnosed Heart Disease in Medicare Patients

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

Dashboard #1


Model Selection: Lasso40
Gender: ☑ (All) ☑ female ☑ male
Ethnicity: ☑ (All) ☑ hispanic or latino ☑ not hispanic or latino
Race: ☑ (All) ☑ black ☑ unknown ☑ white
AgeGroups: □ (All) ☑ Ages 65-74 ☑ Ages 75-84 ☑ Ages 85+ □ Under 65 or Missing

Dashboard #2

## Dashboard #1: Undiagnosed Overview by Selected Population & Factors

This dashboard is a complete overview of all the possible independent factors/variables that can be selected. The map can be filtered by model, ethnicity, race, gender, age, and all risk factors (diabetes, high blood pressure, obesity, statin, aspirin, and other chronic diseases). The bar charts at the bottom ranks US States by undiagnosed percent of selected population or the total count. The percentage for this dashboard is calculated by whichever factors are selected divided by total population of factors selected. For example, 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. If a researcher wanted to know the percentage undiagnosed of a particular selected cohort, they could use the "Undiagnosed Overview by Selected Population & Factors" dashboard. This can lead the research team to formulate a plan of action in getting those at risk prioritized health screening for diagnosing heart disease.

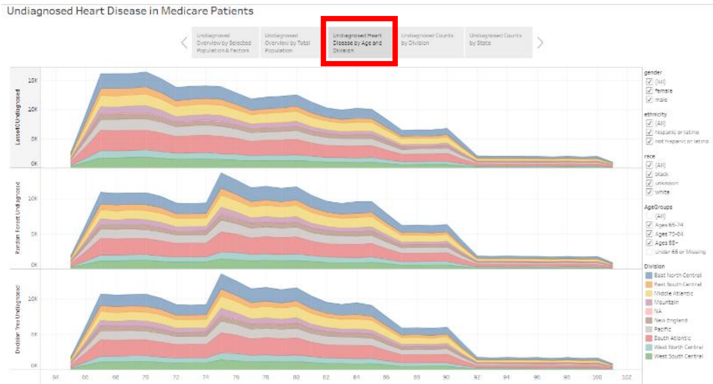
## Dashboard #2: Undiagnosed Overview by Total Population

This dashboard is similar to dashboard #1, except the percentage is calculated by whatever factors are selected divided by the total number of Medicare patients in each US State. 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.

## Supplementary Dashboards

Dashboard #3: Undiagnosed Heart Disease by Age and Division is a line chart of undiagnosed heart disease population by age and our best models (Logistic Regression with 40% threshold with Lasso, Random Forest, and Decision Tree) across the ten different geographic divisions with the ability to filter by cohorts. Dashboard #4: Undiagnosed Counts by Division displays a chart of the undiagnosed percentage of each region and division by each model with the ability to filter by cohorts. Dashboard #5: Undiagnosed Counts by State breaks down Dashboard #4 further into the undiagnosed percentage by each US State. The purpose of these supplementary screens is to help researchers gain a different view of this data based on questions they want to ask and building new ones that don't already exist should be a trivial effort assuming the data is present.

Dashboard #3


Undiagnosed Heart Disease in Medicare Patients

## Conclusion and Discussion

As with any data, there are limitations such as data errors, missing data, or issues in data collection. For example, some of the tables in the database are completely empty. We also acknowledge the source data of Medicare claims data from 2008-2010 is anonymized and partially synthetic and as such it has a limited history for understanding the development of patient journeys for long-term conditions such as heart disease. Ideally, we would want at least 10 years of data containing claims prior to when patients are 65. Another consideration is model replication – we were only able to work with a static dataset and therefore would like to reassess the model every few years based on new patient information. Despite these challenges, it seems possible that one could generate meaningful conclusions using the methods and processes described in the report with a cut of more recent and complete data. This shows that this may be a viable approach of using factors from claims data to potentially identify undiagnosed patients with heart disease.

A potential use of for our experiments is to estimate healthcare costs. Since our project models aim to catch heart disease sooner and identify potential patients that are undiagnosed based on patient journey data, we believe our models could possibly help slow the growth of cost increases.

We experimented with a variety of modeling techniques – some, like Random Forests and Logistic Regression – were found more suitable for our purpose compared to others. While Random Forests are a power modeling technique, we want to be mindful that Logistic Regression is often easier to interpret and therefore might be a better practical choice to aid the researchers that we expect to benefit from our tool (if researchers are uncomfortable at how the model arrives at predictions, they might have hesitations or ethical concerns about using it). Another key piece of our analysis is cost, and we were able to best explore this using different thresholds in the logistic regression analysis. Some model prediction inaccuracies are more costly than others, and we chose to lower the classification threshold for CVD to reflect this.

The Tableau dashboard offers the opportunity to visualize the insights we're able to pull from the data. In our visualization of CVD by age and regional division, we can clearly see that underdiagnoses fall as age increases; we hypothesize this is because people are generally under closer medical supervision as they age, and CVD is more likely to be diagnosed. It's also interesting to see by region/division that the South Atlantic has the most underdiagnosis of CVD. This roll-up view helps focus in on which areas of the country show the most potential for researchers to target in diagnosing untreated CVD.

Our effort to identify and visualize undiagnosed patients from our source dataset falls under a wider goal to contribute to secondary prevention efforts in the medical field. From our literature review, we saw a common theme that secondary prevention is difficult, yet critical to catching CVD in its early stages. We believe analysis and tools such as those outlined in this report will provide new perspective that researchers in the medical field are looking for to make measurable progress toward earlier diagnosis of CVD patients.

All team members have contributed a similar amount of effort.

# References

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE, 14(5). https://doi.org/10.1371/journal.pone.0213653

Alm-Roijer, C., Stagmo, M., Udén, G., & Erhardt, L. (2004). Better Knowledge Improves Adherence to Lifestyle Changes and Medication in Patients with Coronary Heart Disease. European Journal of Cardiovascular Nursing, 3(4), 321–330. https://doi.org/10.1016/j.ejcnurse.2004.05.002

Ariza, M. A., Vimalananda, V. G., & Rosenzweig, J. L. (2010). The economic consequences of diabetes and cardiovascular disease in the United States. Reviews in Endocrine and Metabolic Disorders, 11(1), 1–10. https://doi.org/10.1007/s11154-010-9128-2

Bai, T., Zhang, S., Egleston, B. L., & Vucetic, S. (2018). Interpretable representation learning for healthcare via capturing disease progression through time. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3219819.3219904

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … Virani, S. S. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. Circulation, 139(10). https://doi.org/10.1161/cir.0000000000000659\

Choi, D.-J., Park, J. J., Ali, T., & Lee, S. (2020). Artificial Intelligence for the diagnosis of heart failure. Npj Digital Medicine, 3(1). https://doi.org/10.1038/s41746-020-0261-3

D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., & Kannel, W.B. (2008). General Cardiovascular Risk Profile for Use in Primary Care. American Heart Association, 117(6), 743-753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Fuller, D., Neudorf, J., Lockhart, S., Plante, C., Roberts, H., Bandara, T., & Neudorf, C. (2019). Individual- and area-level socioeconomic inequalities in diabetes mellitus in Saskatchewan between 2007 and 2012: a cross-sectional analysis. CMAJ open, 7(1), E33–E39. https://doi.org/10.9778/cmajo.20180042

Glynn R.J., & Rosner, B. (2005). Comparison of Risk Factors for the Competing Risks of Coronary Heart Disease, Stroke, and Venous Thromboembolism. American Journal of Epidemiology, 162(10), 975–982. https://doi.org/10.1093/aje/kwi309

Johnson, K. W., Soto, J. T., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018, June 5). Artificial Intelligence in cardiology. Journal of the American College of Cardiology. Retrieved February 25, 2022, from https://www.sciencedirect.com/science/article/pii/S0735109718344085?via%3Dihub

Karunathilake, S. P., & Ganegoda, G. U. (2018). Secondary Prevention of Cardiovascular Diseases and application of technology for early diagnosis. BioMed Research International, 2018, 1–9. https://doi.org/10.1155/2018/5767864

Khatibzadeh, S., Farzadfar F., Oliver J., Ezzati M., & Moran A. (2013). Worldwide risk factors for heart failure: A systematic review and pooled analysis. International Journal of Cardiology, 168(2), 1186-1194 https://doi.org/10.1016/j.ijcard.2012.11.065

Khot, U.N., Khot, M.B., Bajzer, C.T., et al. (2003). Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease. JAMA, 290(7), 898–904. https://doi.org/10.1001/jama.290.7.898

Lida, M., Ueda, K., Okayama, A., Ket al. (2003). Impact of elevated blood pressure on mortality from all causes, cardiovascular diseases, heart disease and stroke among Japanese: 14 year follow-up of randomly selected population from Japanese -- Nippon data 80. Journal of Human Hypertension, 7(12) 851-857. https://doi.org/10.1038/sj.jhh.1001602

Müller-Riemenschneider, F., Meinhard, C., Damm, K., Vauth, C., Bockelbrink, A., Greiner, W., & Willich, S. N. (2010). Effectiveness of nonpharmacological secondary prevention of coronary heart disease. European Journal of Cardiovascular Prevention & Rehabilitation, 17(6), 688–700. https://doi.org/10.1097/hjr.0b013e32833a1c95

Papez, V., Moinat, M., Payralbe, S., Asselbergs, F. W., Lumbers, R. T., Hemingway, H., Dobson, R., & Denaxas, S. (2021). Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: A case study in heart failure. JAMIA Open, 4(3). https://doi.org/10.1093/jamiaopen/ooab001

Parthiban, G., & K. Srivatsa, S. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems, 3(7), 25–30. https://doi.org/10.5120/ijais12-450593

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A. & et al. (2022). Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association, 145(8), e152-e639. https://doi.org/10.1161/CIR.0000000000001052

Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., & Kannel, W.B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association, 97(18), 1837-1847. https://doi.org/10.1161/01.CIR.97.18.1837

OMOP Common Data Model - https://www.ohdsi.org/data-standardization/the-common-data-model/

Datasets:

-- condition_era

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.1.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.2.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.3.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.4.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.5.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/condition_era.csv.6.lzo

-- person

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/person.5.2.csv.lzo

-- concept

http://s3.amazonaws.com/ohdsi-sample-data/vocab/CONCEPT.csv.bz2

-- drug_era

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.1.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.2.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.3.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.4.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.5.lzo

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/drug_era.csv.6.lzo

-- concept_ancestor

http://s3.amazonaws.com/ohdsi-sample-data/vocab/CONCEPT_ANCESTOR.csv.bz2

-- location

http://s3.amazonaws.com/ohdsi-sample-data/cmsdesynpuf2.3/location.5.2.csv.lzo