Insurance Insights: Knowledge Discovery Applications for Medicare and Medicaid Claims

Brendan Scott Callender

Martin Solomon Hsu

Andrew Davis Kerr

Sophia Pylin Chung

Data Context

CMS

CENTERS FOR MEDICARE & MEDICAID SERVICES

- Center for Medicare and Medicaid (CMS)
- Synthetic Claims Data (DE-SynPUF) 2008-2010
  - 20 subsets
  - 2 million patients per sample
  - 1 million inpatient claims per sample
- Disclaimer: Limited inferential value

Example Observations

Patient Data

Patient ID	Birth Date	Sex	Race	State	…	Alzheimer’s	…	Stroke
00013D2EFD8E45D1	1923/05/01	1 (Male)	1	26	…	0	…	1

Inpatient Claims Data

Patient ID	Claim ID	Claim From Date	Claim To Date	Claim Amount	…
00013D2EFD8E45D1	196661176988405	2010/03/12	2010/03/13	4000	…

Definitions

- **Medicare/Medicaid** – Federal insurance and welfare programs for (primarily) aged/disabled individuals

Definitions

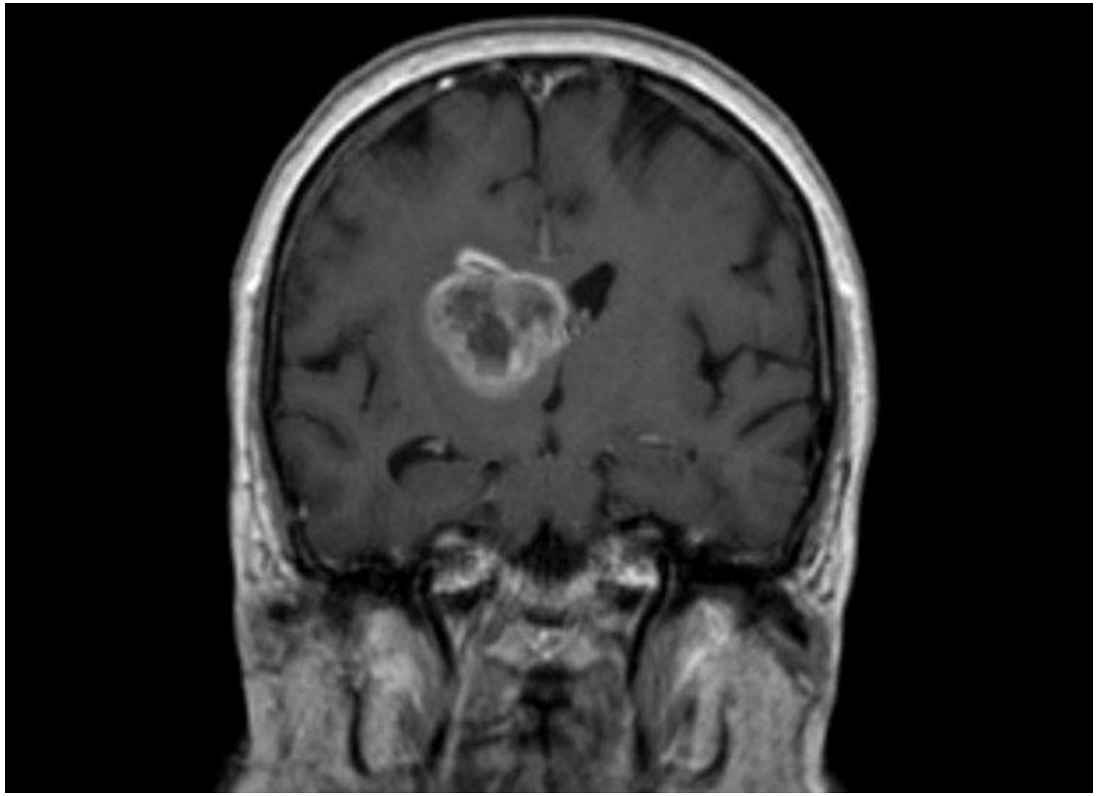
- **Inpatient** – Procedure with overnight patient stays (vs. outpatient)
  - Loose definition

Definitions

- **Claim** – Request for service reimbursement/payment submitted by doctor or patient to insurance company
  - Contains information on patient and procedure including demographics, diagnosis, procedure type, date, cost, etc.

Definitions

- **Chronic condition** – Long-lasting and/or lifelong condition that requires recurring and/or coordinated care
  - E.g. cancer, diabetes, Alzheimer’s, etc.



Project Objectives

1. **Association rules mining:** What **combinations of chronic conditions** (multi-morbidities) tend to **appear together** in a patient with inpatient claims?
2. **Clustering:** What distinct **types of patients** appear in inpatient claims data?
3. **Classification:** Can we **predict** whether a patient has the chronic conditions **diabetes, depression, or heart disease** (or some combination of the three)?

Association Rules Mining

- **Skyline frequent itemsets** – Which chronic conditions appear together most often?
  - Minimum support = 0.10

- **Association rules** – Which chronic conditions do we see most often given the existence of other frequent chronic conditions?
  - Minimum confidence = 0.75

Association Rules Mining

Chronic conditions:

1. Alzheimer's or related disorders or senile
2. Heart Failure
3. (Chronic) Kidney Disease
4. Cancer
5. Chronic Obstructive Pulmonary Disease (COPD)
6. Depression
7. Diabetes
8. (Ischemic) Heart Disease
9. Osteoporosis
10. Rheumatoid Arthritis and Osteoarthritis
11. Stroke/transient Ischemic Attack

Association Rules Mining

Top 3 skyline frequent itemsets:

1. Depression, diabetes, heart disease
2. Heart failure, depression, heart disease
3. Heart failure, kidney disease, heart disease, diabetes

Top 3 association rules:

1. Heart failure, kidney disease → heart disease
2. Heart failure, COPD

   → heart disease
3. COPD, diabetes

   → heart disease

Clustering

Cluster variables

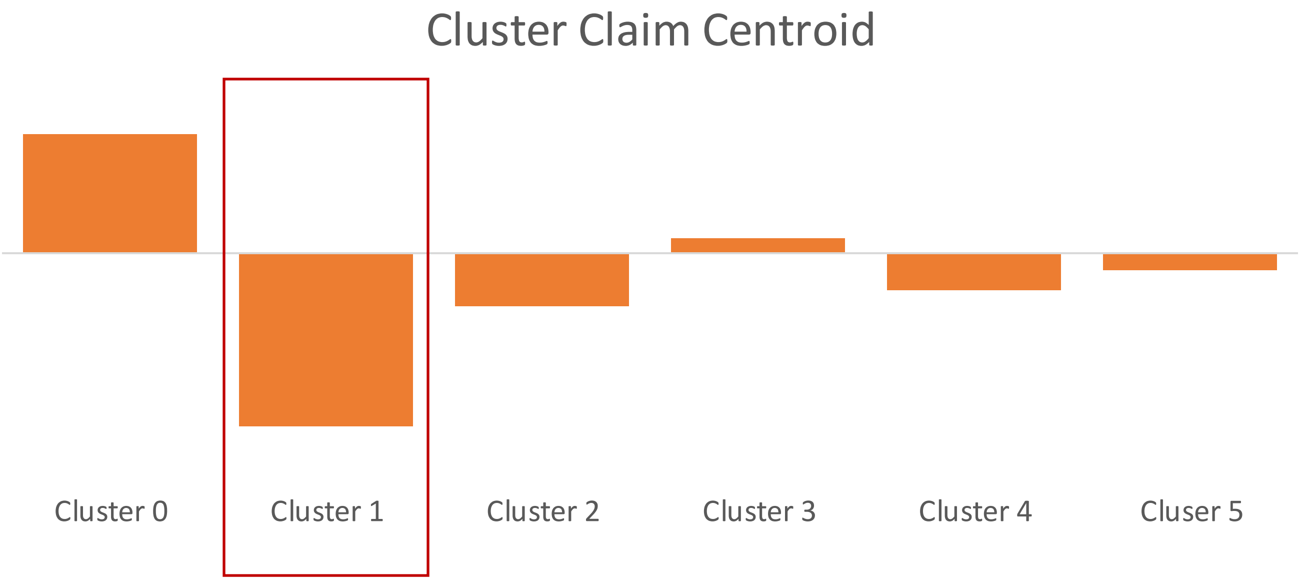
- Age
- Claim payment amount
- Sex
- Chronic condition dummy variables
- Race dummy variables

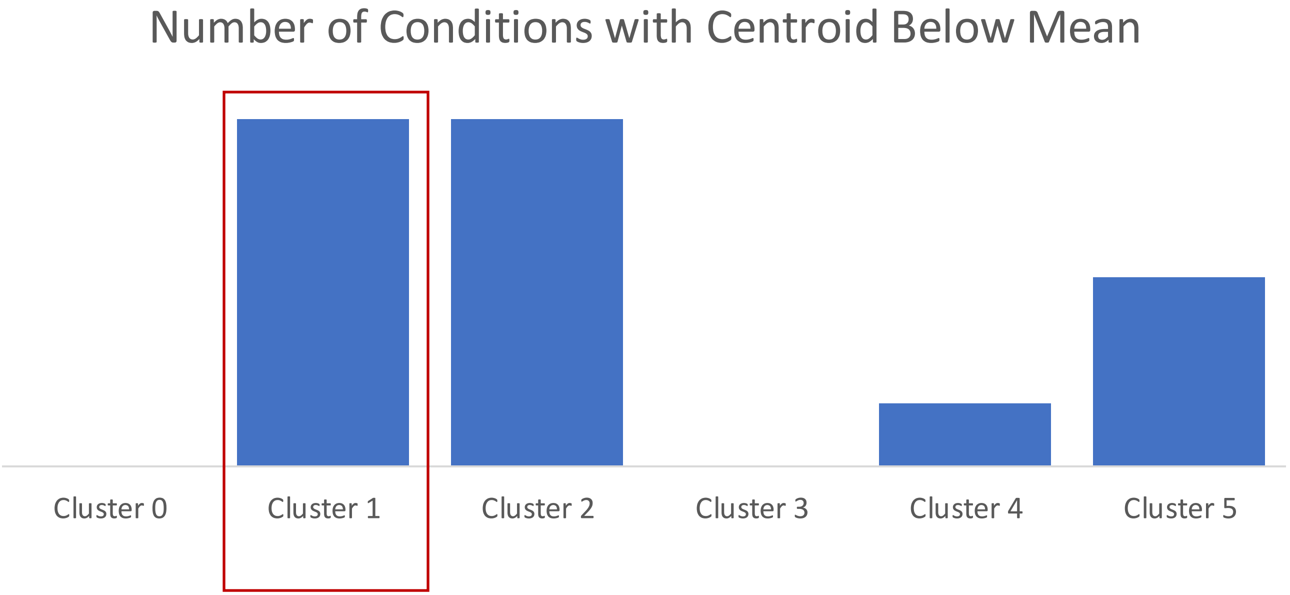
- **K-means**: 6 clusters
  - 2 distinct clusters of interest
  - 4 less distinct clusters

- **DBSCAN:** 6 clusters + outliers
  - 10% random sample of dataset
  - Most clusters trend young, female
  - 3 distinct clusters of interest
  - Outliers trend older, expensive

Clustering

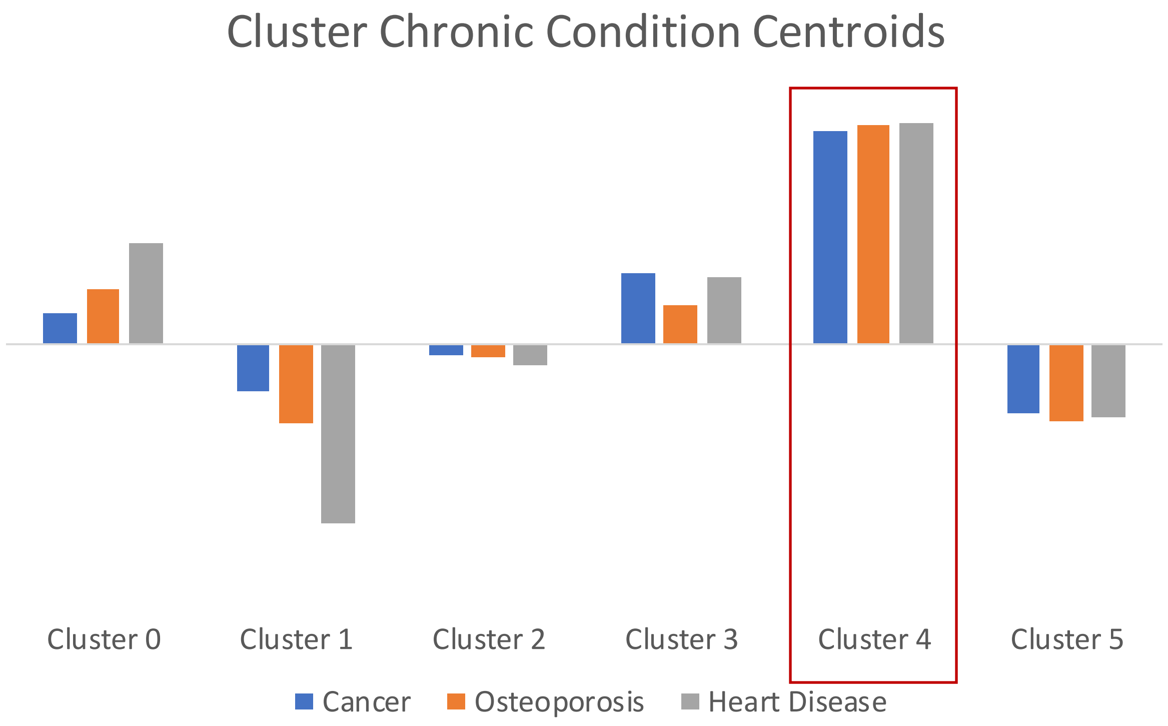
- K-means cluster 1:
  - Lower claim counts
  - Lower claim amounts
  - Less chronic conditions

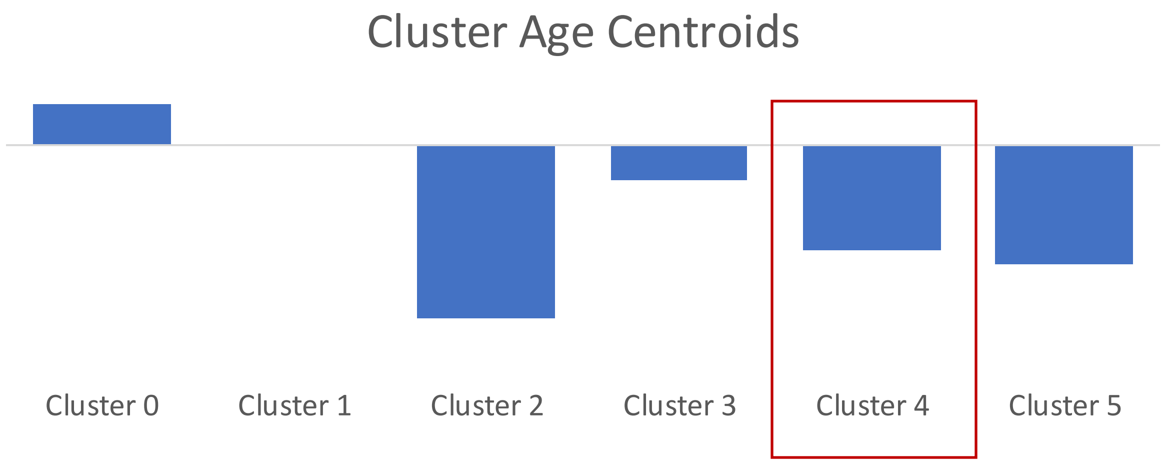




Clustering

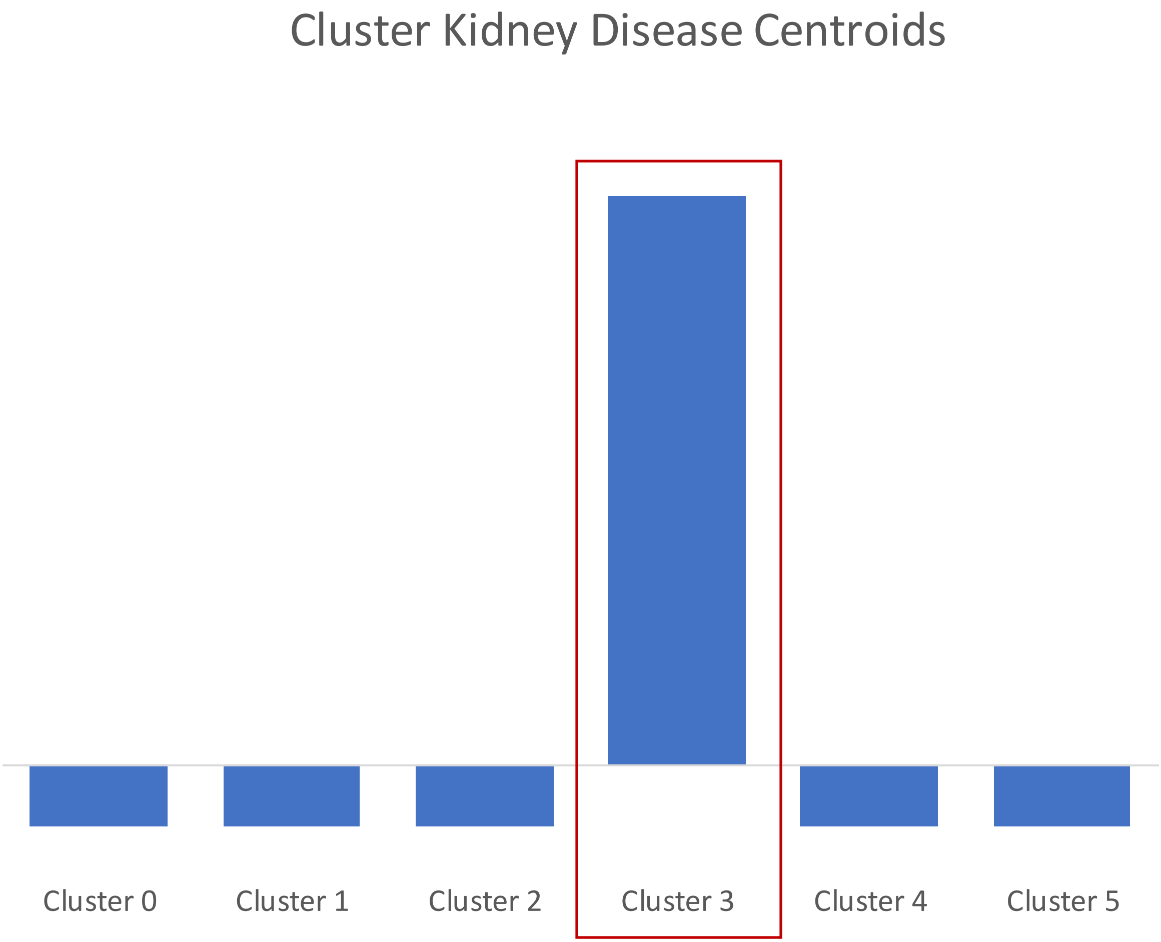
- K-means cluster 4:
  - High rates of cancer, COPD, osteoporosis, heart disease
  - Younger males





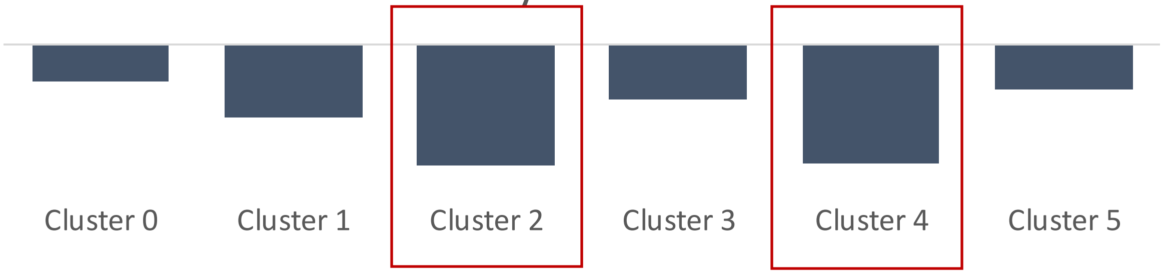
Clustering

- DBSCAN cluster 3:
  - High levels of chronic kidney disease

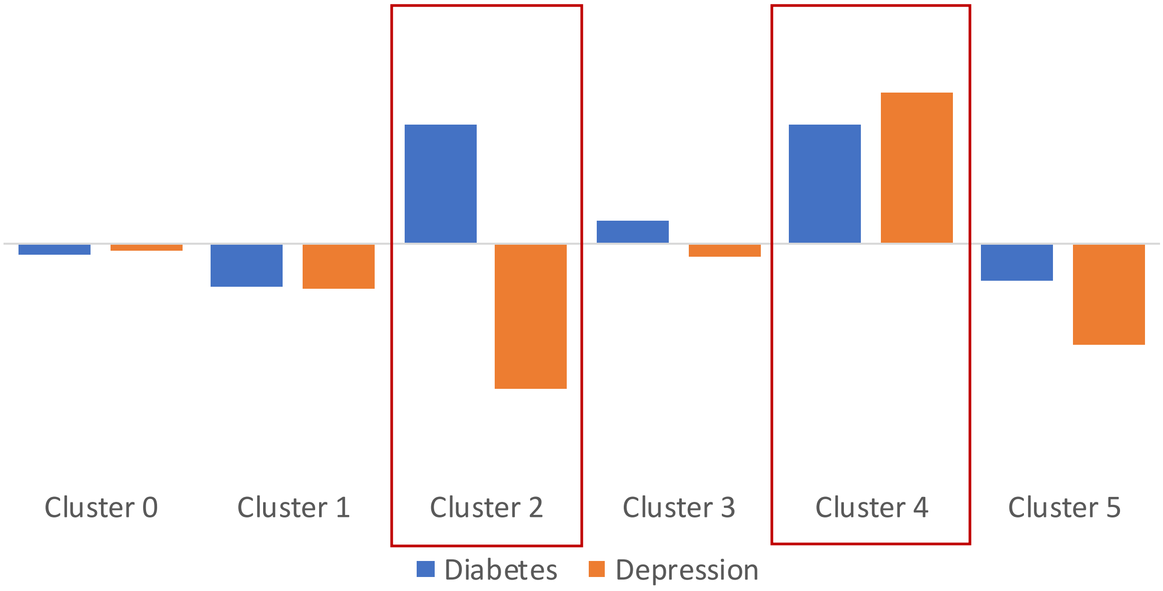


Clustering

- DBSCAN clusters 2 & 4:
  - Lower payment amounts
  - Lower rates of most chronic disease
  - High rates of diabetes
  - High number of claims
  - Very low rates of depression in 2, very high rates of depression in 4



Cluster Payment Centroids



Cluster Diabetes & Depression Centroids

Classification

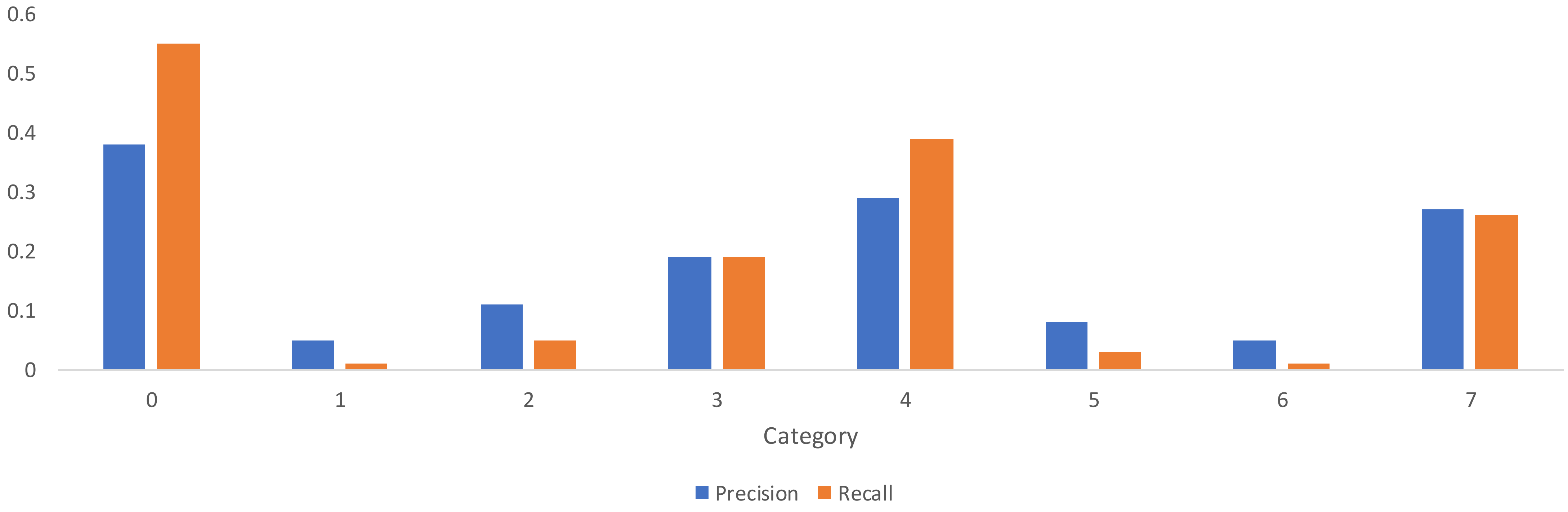
- Two algorithms:
  - **K-Nearest Neighbors** – 10 neighbors
  - **Random Forest** – 5 attributes and 1000 patients per tree, 5 trees, 0.02 gains threshold
- Predict **depression, diabetes, heart disease**
- Performance
  - Relatively poor accuracy
  - High runtime

8 Categories:

- 0: None of 3 diseases
- 1-3: 1 of 3 diseases
  - *1: Depression only*
  - *2: Diabetes only*
  - *3: Heart disease only*
- 4-6: 2 of 3 diseases
  - *4: Diabetes, heart disease*
  - *5: Depression, heart disease*
  - *6: Depression, diabetes*
- 7: All 3 diseases

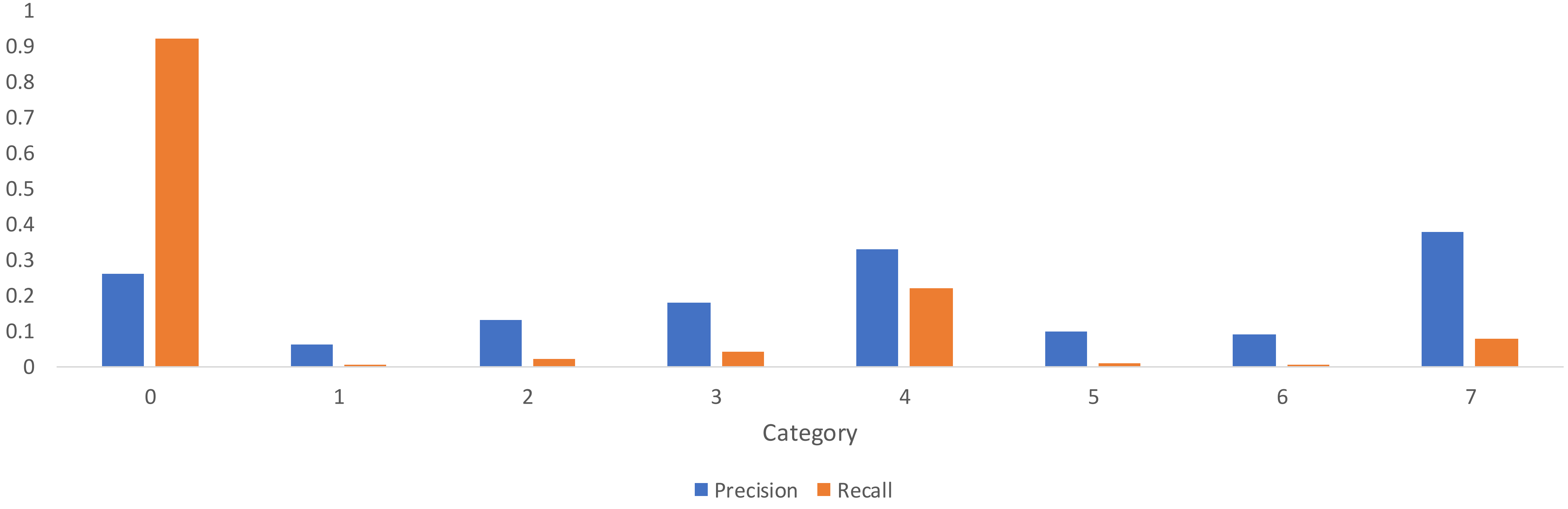
Classification

10-nearest neighbors (Accuracy: 28.2%)



Classification

Random Forest (Accuracy: 26.8%)



Ethical Concerns

- Age, race, sex in analysis
  - Included due to limited information in synthetic dataset
  - May lead to a system that does not demonstrate equality or group parity in analysis

Conclusions

- Association rule mining most effective
- Classification by far the worst
  - Didn’t have access to the best predictors for the task
- Clustering mixed results
  - Found some insightful and distinct clusters
  - Try larger k's?

Questions?