Insurance Insights:
Knowledge Discovery Applications for Medicare and Medicaid Claims

Brendan Callender, Martin Hsu, Andrew Kerr, Sophia Chung

ABSTRACT. *The following report analyzes and discusses the use of data mining, clustering, and classification on DE-SynPUF data from the Center for Medicare & Medicaid Services. The data involves inpatient beneficiaries, with specific focus on demographics, chronic conditions, and claim payments. We begin by investigating sets of chronic conditions that individual patients tend to have using frequent itemsets and association rules. This is followed by a comparison of k-means clustering to DBSCAN in order to understand the characteristics of the data, and is concluded with a comparison of k-nearest neighbors (KNN) to Random Forest in regards to predicting a patient's chronic conditions. The results show a successful identification of six skyline frequent itemsets, seven association rules, and a select few cluster insights. However, the classification models yielded poor results due to data limitations.*

1. INTRODUCTION

The Center for Medicare & Medicaid Services is a federal agency which oversees Medicare, Medicaid, and other health insurance programs. They have created DE-SynPUF, a synthetic public use file, mimicking patient data from 2008 to 2010. The purpose of the DE-SynPUF is to aid in the creation and development of medical software and allow researchers to practice using medical data. For this project, we utilize the data to answer three analytical questions requiring the use of three different knowledge discovery techniques.

2. DATA

2.1 Data Description

Our data pertains to inpatient beneficiaries, referring to patients who stay in a hospital overnight while undergoing treatment. The data was originally separated into two tables: one for patients with 32 variables and one for inpatient claims with 81 variables. We focus on variables pertaining to demographics, chronic conditions, and claim payment amounts, where each patient can have one or more claims.

2.2 Data Manipulation

First, we filtered our data to only include patients with inpatient claims. Originally, chronic conditions were coded to 1 if a patient had it and 2 if they did not have it, however we recoded chronic conditions to 0 if a patient did not have it and 1 if they had it. Similarly, we recoded sex to 0 being female and 1 being male where sex was originally coded as 1 being male and 2 being female.

Additionally we engineered two features: estimated age in days and claim count. Estimated age was calculated by finding the difference in days between a patient's date of birth and the 31st, of December, of the first year a patient appeared in the data. Claim count is a patient's number of claims.

3. RESEARCH QUESTIONS

3.1 Association Rules

According to the National Library of Medicine, "Multimorbidity is the coexistence of two or more chronic diseases in an individual."[1] Using the patient data, we investigated sets of chronic conditions that tend to appear in patients experiencing multimorbidity. We answer this question using data mining techniques to uncover frequent sets of chronic conditions shared by different patients, as well as association rules between different conditions.

3.2 Clustering

Our goal for clustering the data was to find distinct groupings of patients that exist within the dataset. Furthermore, we wanted to determine the distinct characteristics of each of these clusters. For instance, we aimed to discover whether there exist large groupings of patients who are much older on average, suffer from several chronic illnesses, and have very high claim payment amounts. In order to uncover different clusters, we used both k-means clustering and DBSCAN.

3.3 Classification

In regard to classification, we sought to investigate whether we could predict if a patient has any combination of depression, diabetes, and ischemic heart disease. Since there are eight resulting possible conditions, we treat this as an eight-class classification problem.

Table 1. Mapping of Conditions to Class

Class	Conditions
0	None of the three diseases
1	Depression
2	Diabetes
3	Ischemic Heart Disease
4	Diabetes, Ischemic Heart Disease
5	Depression, Ischemic Heart Disease
6	Depression, Diabetes
7	All three diseases

In order to predict the condition of a patient, we used indicators for other chronic illnesses, patient demographic information, and information relating to that patient's history of claims. Some claim related variables include the claim payment amount and the number of claims attributed to that patient. We used two classification methods: Random Forest and KNN.

4. RESULTS & ANALYSIS

4.1 Association Rules

Each multimorbidity condition in *Table 2* is considered a frequent multimorbidity condition because it occurs in at least 10% of patients in the data. An example interpretation for a condition in *Table 2* goes as follows: 14.5% of all patients have depression, diabetes, and ischemic heart disease.

[1] Smith, S. M., & O'Dowd, T. (2007, April). Chronic diseases: What happens when they come in multiples?

Table 2. Skyline Frequent Itemsets (minSup = 0.1)

Multimorbidity Condition	Support
Depression, Diabetes, Ischemic Heart Disease	0.145
Heart Failure, Depression, Ischemic Heart Disease	0.131
Heart Failure, Chronic Kidney Disease, Ischemic Heart Disease, Diabetes	0.121
RA/OA, Ischemic Heart Disease	0.121
Chronic Obstructive Pulmonary Disease, Diabetes, Ischemic Heart Disease	0.109
Heart Failure, Chronic Obstructive Pulmonary Disease, Ischemic Heart Disease	0.108

We determined association rules with a minimum confidence of 75%. An example interpretation for the first row in *Table 3* goes as follows: 83.4% of patients with heart failure, chronic kidney disease, and diabetes also have ischemic heart disease.

Table 3. Singleton Association Rules (minConf = 0.75)

Singleton Association Rule	Confidence
Heart Failure, Chronic Kidney Disease, Diabetes → Ischemic Heart Disease	0.834
Heart Failure, Chronic Obstructive Pulmonary Disease → Ischemic Heart Disease	0.825
Chronic Obstructive Pulmonary Disease, Diabetes → Ischemic Heart Disease	0.818

Continued from Table 3

Heart Failure, Depression → Ischemic Heart Disease	0.792
Heart Failure, Chronic Kidney Disease, Ischemic Heart Disease → Diabetes	0.785
Depression, Diabetes → Ischemic Heart Disease	0.778
RA/OA → Ischemic Heart Disease	0.754

4.2 Clustering

Our best k-means clustering output was using k=6 clusters. Out of the clusters shown in *Table 4*, clusters 1 and 5 contain the most patients while cluster 4 contains the least. Analyzing the centroid values for each cluster led to deeper insights into clusters 1 and 4.[2] Cluster 1 consists of patients with lower claims counts, claim payments, and less chronic conditions overall. Meanwhile, cluster 4 consists of younger male patients who tend to have cancer, chronic obstructive pulmonary disease, osteoporosis, and ischemic heart disease.

Table 4. k-means Clustering (standardized metrics)

Cluster	Size	Avg Dist. to Center	SSE
0	2044	4.117	36431.38
1	14145	2.758	116736.16
2	4594	4.455	96954.64
3	3118	4.159	57115.50
4	743	3.836	11670.79
5	13136	3.616	176695.85

[2] See Table B in Appendix.

Due to computational and time restraints, we ran DBSCAN using a 10% sample, or 3778 patients. We define a core point as a point with a minimum of 18 neighbors within a radius of 3. This resulted in 6 clusters and 823 outliers. As seen in *Table 5*, cluster 0 contained a majority of patients. On the other hand, clusters 2, 4, and 5 barely contained any. Again, we analyze the centroid values for deeper insights.[3] Because all the clusters were centered at below average age, we deduce that older patients tended to be classified as outliers. The same logic can be applied to claim payment amount and sex: patients with higher claim payment amounts tended to be classified as outliers because all the clusters were centered at below average claim payment amounts, and male patients tended to be classified as outliers because all clusters were female dominant. Cluster 1 has the most white people, cluster 2 has patients who tend to have more claims, cluster 3 has patients who tend to have kidney disease, cluster 4 has patients who tend to have depression, and cluster 5 has the most black people. Additionally, clusters 2 and 4 share the same standardized values for all chronic illnesses except depression, in which case cluster 2 is significantly below average and cluster 4 significantly above average.

Table 5. DBSCAN Clustering (standardized metrics)

Cluster	Size	Avg Dist. to Center	SSE
0	2527	3.351	29419.53
1	197	2.798	1614.82
2	45	1.093	73.91
3	126	2.772	999.5
4	33	0.902	34.17
5	27	2.235	140.9

4.3 Classification

The best KNN model was run using k=10, while the best Random Forest hyperparameters were 5 attributes per tree, a training sample size of 1000 patients per tree, 5 trees, and a gains threshold of 0.02. Both KNN and Random Forest classification methods resulted in poor accuracies as seen in *Table 6*. However, we would expect an overall accuracy of 12.5% with random guessing, so the two models are able to more than double that success.

Table 6. Accuracy Measurements

Method	Overall Accuracy
KNN	28.2%
Random Forest	26.8%

The precision metrics for classes 0, 5, and 7 were the best for both KNN and Random Forest, ranging between 0.27 and 0.40.[4]

[3] See Table D in Appendix.

[4] Refer to *Table 1* for class descriptions.

Table 7. Precision Measurements

Class	KNN	Random Forest
0	0.38	0.26
1	0.05	0.06
2	0.11	0.13
3	0.19	0.18
4	0.29	0.33
5	0.08	0.10
6	0.05	0.09
7	0.27	0.38

With regard to recall, the same classes performed the best for KNN, however the recalls for Random Forest are significantly worse.

Table 8. Recall Measurements

Class	KNN	Random Forest
0	0.55	0.92
1	0.01	0.003
2	0.05	0.02
3	0.19	0.04
4	0.39	0.22
5	0.03	0.01
6	0.01	0.001
7	0.26	0.08

5. CONCLUSION

The three knowledge discovery techniques we used to analyze the synthetic claims data provided mixed results. Data mining for frequent multimorbidity conditions and association rules effectively linked different chronic conditions to each other. We uncovered conditions shared by more than 10% of patients and association rules with greater than 75% confidence. These results answered what chronic conditions tend to appear together.[5] Meanwhile, clustering the data in an attempt to find distinct groups of patients gave us lots of results, with only a handful being meaningful. Using k-means clustering and DBSCAN, we were able to uncover several unique clusters of patients.[6] Lastely, our attempts to classify patients' conditions produced poor results. When predicting whether a patient had any combination of depression, heart disease, and diabetes, we used both KNN as well as Random Forest. These methods produced accuracies of 28.2% and 26.8% respectively.[7] Ultimately, we were unable to confidently classify patients into one of the eight possible sets of conditions.

The success of data mining for frequent itemsets and association rules can be attributed to the structure of the data being suited to this kind of analysis. Since the data contained indicators for each patient containing information for whether or not they have a certain chronic condition, the analysis was straightforward and effective to carry out.

[5] See *Table 2* and *Table 3* for a complete description of the results.
[6] See Section 4.2.
[7] See Section 4.3 for additional metrics.

On the other hand, performing clustering and classification on the data was much more challenging. When clustering the data, we ran into issues with there not being many meaningful variables to include when clustering, besides the variables for different chronic conditions. This analysis would likely have been more effective if we had access to more information about each patient and their stay at the hospital. Similarly, our classification models struggled to effectively predict patients' conditions due to the lack of meaningful predictors. The most meaningful predictors we had access to were, again, the information relating to other chronic conditions. With more information regarding each patient and their condition, we believe we could improve the results of the classification models drastically.

6. APPENDIX

A. 6-Means Clustering Standardized Intercluster Distances

	1	2	3	4	5
0	4.916	5.678	5.153	8.675	4.383
1	---	4.181	4.281	7.853	2.407
2	---	---	5.199	7.762	4.208
3	---	---	---	8.377	3.512
4	---	---	---	---	7.811

B. 6-Means Clustering Standardized Cluster Centers

Feature	0	1	2	3	4	5
Est Age in Days	0.077	0.004	-0.317	-0.063	-0.193	-0.219
CLM_PMT_AMT	0.129	-0.252	-0.115	-0.044	-0.102	-0.227
Claim Count	0.219	-0.319	-0.098	0.029	-0.067	-0.032
BENE_SEX_IDENT_CD	-0.003	0.015	-0.012	0.053	0.322	0.052
SP_ALZHDMTA	0.363	-0.324	-0.0370	0.112	0.401	0.441
SP_CHF	0.366	-0.578	-0.066	0.204	0.319	0.282
SP_CHRNKIDN	0.401	-0.455	-0.053	3.057	-0.322	-0.327
SP_CNCR	0.101	-0.148	-0.035	0.229	0.679	-0.216
SP_COPD	0.253	-0.331	-0.063	0.019	0.674	-0.255
SP_DEPRESSN	0.205	-0.315	-0.031	0.165	0.471	0.470
SP_DIABETES	0.370	-0.631	-0.060	0.148	0.490	0.445
SP_ISCHMCHT	0.326	-0.566	-0.066	0.216	0.704	-0.226
SP_OSTEOPRS	0.175	-0.246	-0.041	0.124	0.697	-0.243
SP_RA_OA	0.178	-0.255	-0.035	0.036	0.306	-0.073
SP_STRKETIA	4.061	-0.246	-0.240	0.082	-0.414	0.435
BENE_RACE_CD_1	0.081	0.424	-2.359	-0.007	0.391	-0.343
BENE_RACE_CD_2	0.007	-0.332	2.158	-0.042	0.284	-0.188
BENE_RACE_CD_3	-0.064	-0.185	1.238	-0.143	-0.143	-0.143
BENE_RACE_CD_5	-0.141	-0.142	-0.142	-0.063	-0.193	-0.219

C. DBSCAN (Core Point Radius of 3 with 18 Minimum Neighbors) Standardized Intercluster Distances

	1	2	3	4	5
0	4.524	2.412	3.454	2.290	6.736
1	---	4.855	5.612	4.903	6.879
2	---	---	3.998	2.916	6.740
3	---	---	---	3.827	7.472
4	---	---	---	---	6.971

D. DBSCAN Standardized Cluster Centers

Feature	0	1	2	3	4	5
Est Age in Days	-0.121	-0.235	-0.316	-0.229	-0.328	-0.362
CLM_PMT_AMT	-0.137	-0.272	-0.447	-0.208	-0.443	-0.167
Claim Count	-0.019	-0.119	1.165	-0.232	-0.858	0.266
BENE_SEX_IDENT_CD	-0.034	-0.327	-0.640	-0.465	-0.640	-0.558
SP_ALZHDMTA	-0.020	-0.381	-0.921	0.003	-0.921	-0.624
SP_CHF	-0.050	-0.387	-0.634	-0.318	-0.634	-0.634
SP_CHRNKIDN	-0.330	-0.330	-0.330	3.030	-0.330	-0.330
SP_CNCR	-0.084	-0.413	-0.501	-0.224	-0.501	-0.501
SP_COPD	-0.027	-0.242	-0.652	-0.375	-0.652	-0.571
SP_DEPRESSN	-0.058	-0.305	-0.985	-0.096	1.015	-0.689
SP_DIABETES	-0.075	-0.297	0.804	0.154	0.804	-0.257
SP_ISCHMCHT	-0.044	-0.351	-0.503	-0.226	-0.503	-0.503
SP_OSTEOPRS	-0.062	-0.339	-0.448	-0.363	-0.448	-0.348
SP_RA_OA	-0.246	-0.246	-0.246	-0.246	-0.246	-0.246
SP_STRKETIA	0.414	-2.415	0.414	0.414	0.414	-2.415
BENE_RACE_CD_1	-0.325	3.071	-0.325	-0.325	-0.325	-0.325
BENE_RACE_CD_2	-0.174	-0.174	-0.174	-0.174	-0.174	5.747
BENE_RACE_CD_3	-0.147	-0.147	-0.147	-0.147	-0.147	-0.147
BENE_RACE_CD_5	-0.121	-0.235	-0.318	-0.229	-0.328	-0.362

E. 10-Nearest Neighbors Classification Confusion Matrix

	Observed							
Predicted	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	4441	712	1062	2204	1534	624	369	704
1	107	21	37	88	64	29	12	39
2	303	56	185	319	378	101	83	221
3	1342	335	641	1217	1315	451	257	687
4	1344	393	1080	1803	3291	823	547	2219
5	137	31	78	160	174	68	37	129
6	36	17	30	46	71	24	16	51
7	389	175	392	673	1573	385	236	1414

F. Random Forest (5 Attributes and Training Sample Size 1000 per Tree, 5 Trees, Gains Threshold 0.02) Classification Confusion Matrix

	Observed							
Predicted	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	7456	1500	2749	5313	5555	1890	1147	3104
1	2	5	10	9	25	3	5	19
2	60	30	69	98	124	39	21	98
3	142	44	126	249	412	111	56	267
4	364	141	450	687	1879	372	258	1458
5	21	4	9	27	46	18	4	44
6	3	0	1	3	3	1	2	10
7	51	16	91	124	356	71	64	464

7. REFERENCES

CMS 2008-2010 Data Entrepreneurs' synthetic public use file (DE-SynPUF). CMS.gov. (n.d.). https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-claims-synthetic-public-use-files/cms-2008-2010-data-entrepreneurs-synthetic-public-use-file-de-synpuf

Smith, S. M., & O'Dowd, T. (2007, April). *Chronic diseases: What happens when they come in multiples?*. The British journal of general practice : the journal of the Royal College of General Practitioners. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2043326/#:~:text=Multimorbidity%20is%20the%20coexistence%20of,the%20age%20of%2065%20years.