Palliative Care Diagnosis Clustering and Dimensionality Reduction

The purpose of the analyses in this report was to determine if there were groupings of diagnoses in the Electronic Health Records of patients who were assigned to Palliative Care. Palliative care is health care that is given to patients with incurable terminal illnesses to make their existence more comfortable. This is contrasted with normal health care which is focused on curing patients.

The ultimate goal is to use the results from these analyses to find matching patients in the larger patient population. For example, if it were determined that palliative care patients often had both heart disease and chronic pulmonary disease, then matching patients would be located that had both heart disease and chronic pulmonary disease. The two groups, the Palliative Care patients, and the Matching patients would then be compared to determine whether there were significant differences in outcomes.

Two types of machine learning algorithms were used. KMeans was used to determine the effects of diagnosis clustering. Principal components analysis (PCA) was used to determine the effects of dimension reduction.

Both methods have plusses and minuses.

Principal Investigator

These analyses were performed by Thomas Arnold, PhD, who is a Senior Data Scientist in Population Health. Dr. Arnold performed all actions, which included pulling the data, cleaning the data, and performing the analyses. This is one part of a larger project.

Data Description

The data for these analyses consisted of 259 dichotomous variables which indicated the Condition Classification System (CCS) codes for each of the 11,332 palliative care patients. The CCS coding system is one of the primary diagnosis groupers used in healthcare. The CCS coding system groups about 70,000 International Classification of Disease Version 10 (ICD10) codes into about 260 diagnosis groups. The CCS diagnosis code grouper provides a much more convenient subset of diagnoses to work with than the original ICD10 code set.

Two CCS codes had been dropped from the initial data download as part of the preliminary data cleaning. One of the CCS codes that had been removed was associated with palliative care and the other was a generic CCS grouper that had no specific meaning. A sample of the data is provided below for the first 10 patients.

Sample Data for 10 Patients with 10 CCS Codes

ID	0	1	2	3	4	5	6	7	8	9
Abdominal hernia	0	0	0	0	0	0	0	0	0	0
Abdominal pain	0	0	0	0	0	0	0	0	0	0
Acquired foot deformities	0	0	0	0	0	0	0	0	0	0
Acute and unspecified renal failure	0	0	0	1	0	0	0	1	0	0
Acute bronchitis	0	0	0	0	0	0	0	0	0	0
Acute cerebrovascular disease	1	0	0	0	0	0	0	0	0	0
Acute myocardial infarction	0	0	0	1	0	0	1	0	0	0
Acute post hemorrhagic anemia	0	0	0	0	0	0	1	0	0	0
Adjustment disorders	0	0	0	0	0	0	0	0	0	0
Administrative/social admission	0	0	0	0	0	0	0	0	1	0

The patient data typically only has a few CCS columns marked as one for each patients with the rest of the CCS columns marked as zero.

Data Exploration

A correlation analysis was created to determine which of the diagnosis groups were most highly correlated. The diagonal values were set to zero. A sample of the first 10 feature correlations is shown below.

Sample Correlations (10x10)

#	Feature	1	2	3	4	5	6	7	8	9	10
1	Abdominal hernia	.000	.033	-.008	.070	.003	.000	.015	.075	.004	.002
2	Abdominal pain	.033	.000	-.009	-.020	-.009	-.035	-.016	-.005	.028	.004
3	Acquired foot deformities	-.008	-.009	.000	.002	-.003	-.003	-.017	.019	.020	.000
4	Acute and unspecified renal failure	.070	-.020	.002	.000	.007	-.030	.230	.165	-.003	-.025
5	Acute bronchitis	.003	-.009	-.003	.007	.000	-.016	.017	-.015	-.007	.005
6	Acute cerebrovascular disease	.000	-.035	-.003	-.030	-.016	.000	.023	.011	.013	-.024
7	Acute myocardial infarction	.015	-.016	-.017	.230	.017	.023	.000	.047	-.008	.016
8	Acute post hemorrhagic anemia	.075	-.005	.019	.165	-.015	.011	.047	.000	.025	.014
9	Adjustment disorders	.004	.028	.020	-.003	-.007	.013	-.008	.025	.000	.025
10	Administrative/social admission	.002	.004	.000	-.025	.005	-.024	.016	.014	.025	.000

The most highly correlated features were examined. These are shown below. Many seem to make logical sense.

Most Highly Correlated Features

Feature 1	Feature 2
Abdominal hernia	Intestinal obstruction without hernia
Abdominal pain	Nausea and vomiting
Acquired foot deformities	Multiple sclerosis
Acute and unspecified renal failure	Fluid and electrolyte disorders
Acute bronchitis	Chronic obstructive pulmonary disease and bronchiectasis
Acute cerebrovascular disease	Paralysis
Acute myocardial infarction	Acute and unspecified renal failure
Acute posthemorrhagic anemia	Gastrointestinal hemorrhage
Adjustment disorders	Benign neoplasm of uterus
Administrative/social admission	Chronic kidney disease

Feature Engineering

I set up a pipeline for feature engineering. The pipeline had two components.

1. Skew removal using Log Transformations
2. Scaling

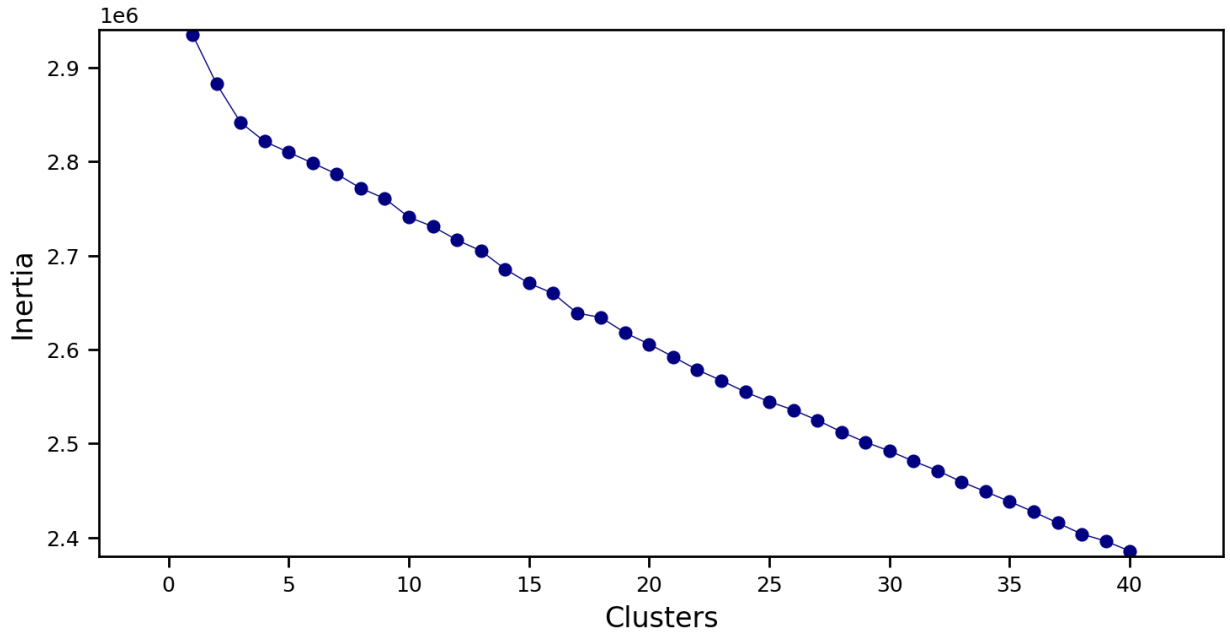
The skew removal process involved taking the log of the features with skew greater than .75. Two types of scalers were tried. The standard scaling was used with the KMeans calculations and the MinMax scaling was used with the PCA calculations.

1. Standard scaling
2. MinMax Scaling

KMeans Diagnosis Clustering

The KMeans cluster analysis was set up with 2 clusters. This provided a "proof of concept." A loop was set up to determine if there was a spot where there was an "elbow" in the inertia based on the number of clusters. There may be a slight elbow at 3 clusters, but it appeared that there was a linear reduction in inertia through 40 clusters.

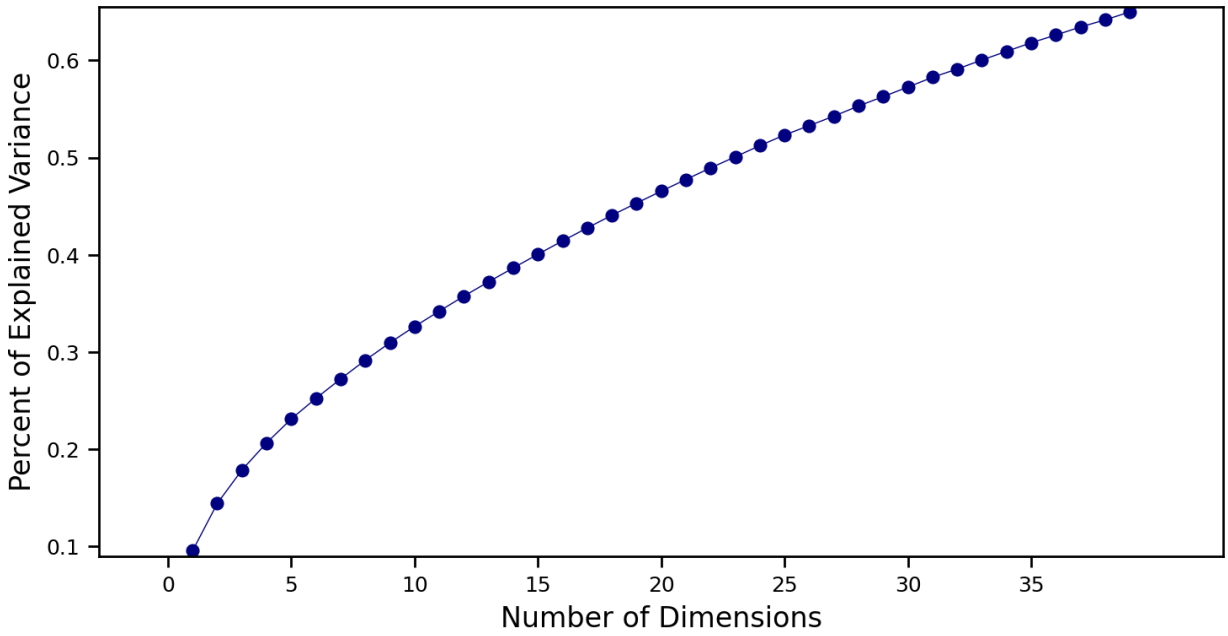
KMeans Inertia by Cluster Count



Principal Components Analysis (PCA) Dimension Reduction

The PCA analysis was set up with 2 dimensions. This provided a "proof of concept." A loop was set up to determine if there was a spot where there was an "elbow" in the percent of explain variance based on the number of dimensions. There may be a slight elbow at 4-5 dimensions, but it appeared that there was a linear increase in the percent of explained dimensions through 40 dimensions.

PCA Percent of Explained Variance by Number of Dimensions



Summary

Both KMeans and PCA provided a method for grouping the CCS diagnoses. Further analyses is needed to determine which method provides an optimal grouping method. Both provide similar results when it comes to numbers of clusters or dimensions. There appeared to be linear improvements in accuracy through 40 groups.

There were several lessons learned. These included the following.

Lessons Learned

1. Correlation matrix
    a. Look at the inter item correlations
    b. Produce top pairs by correlation
2. Feature engineering using Pipelines
    a. Skew reduction
    b. Scaling
        i. Standard scaling
        ii. MinMax scaling
3. Group identification
    a. KMeans Cluster Analysis
    b. Principal Components Analysis

Future Directions

The analyses presented above produced a general approach to the problem of patient diagnosis grouping. It did not appear that there was substantial support for either of these methods in the literature on diagnosis grouping. One of the papers on diagnosis grouping used network analysis and graph theory. This course did not cover either, and so further investigation is required.

I was not able to use either of these methods before taking the course, so just getting the modesl to run seemed a worthwhile goal. I will need to play around with these a bit more to come to a definitive solution.

For example, do the clusters produced by the KMeans cluster analysis match the dimensions created with Principal Components Analysis? How would one compare the models? The KMeans method produces groups, and the PCA method produces factor weights. If I use these models with the rest of the patient data will I be able to find similar groups of patients to match against? I have more questions than answers at this point.