Rohan Agarwal

Medicare Analysis: Identifying Areas of Improvement

Executive Summary

From the US government's Medicare data, k-means clustering was used in order to find HCPCS processes in which support could be improved. The data in consideration was sampled from five different states and detailed by HCPCS process description.

The attributes extracted from the data are the process description, allowed payment, submitted payment, Medicare payment, number of beneficiaries, and number of services. Two custom parameters were calculated, the discount provided by Medicare and the coverage that Medicare provides. In order to gain a representative sample of every process provided by a state, a one-cluster medoid was generated from a subset of the data, broken down based on the process. The parameters used for clustering were the number of beneficiaries, the number of services in a process, the discount, and the coverage, all of which were standardized.

Medoids were generated for all processes in the randomly selected states of Illinois, Virginia, California, New York, and Michigan. Each state had from 2400-3500 medoids representing their processes. After collecting all of the medoids, they were combined and clustered using K-means clustering. There were seven clusters, out of which one cluster contained low-discounted processes, some of which with low coverage as well, that needed to be addressed.

Out of the processes that needed to be addressed, two types of processes were identified: those which need attention across all sample states, and those which require attention in one state but are seemingly better covered in other states. Based on this sample analysis one can identify areas of improvement not only for Medicare as a whole, but also in specific states in comparison to other geographies.

Rohan Agarwal

Introduction

The US government recently published data regarding its utilization of Medicare, specific to individual providers. There are several million records to parse, from which this analysis has dealt with samples for several states. The question at hand to answer is which healthcare processes have the weakest Medicare coverage. Therefore he following analysis uses clustering to identify specific medical processes that are under-covered by Medicare and deserve more attention.

Data Processing

The raw data was downloaded as a text file, from which a database .db file was created using SQLite in Windows. The database was then parsed and analysed in R. The specific attributes that were used in this analysis were the average submitted charges by the provider, average allowed charge by Medicare, average Medicare payment amount, number of distinct Medicare beneficiaries, number of services provided, and the corresponding HCPCS process description for all of these quantities.

From this data, there were two additional attributes calculated:

1) The "discount" given by Medicare to its participants, which is the ratio of the payment that Medicare allows to the amount submitted by the provider subtracted from 1. The lower the Medicare amount is, the higher the discount provided.

$$discount = 1 - \frac{Medicare\ allowed\ amount}{Provider\ submitted\ amount}$$

2) The "coverage" that Medicare provides, which is the ratio of the amount Medicare pays to the amount it allows. The more Medicare pays, the higher the coverage.

$$coverage = \frac{Medicare\ payment\ amount}{Medicare\ allowed\ amount}$$

The final attributes considered for clustering were the number of beneficiaries, number of services, discount, and coverage. Ideally a cluster with a high number of beneficiaries and services with a low discount and coverage would be an area of improvement as Medicare coverage would be lacking but required. Before clustering, the data was pre-processed by standardizing all values with the following formula:

$$y = \frac{y - y_{max}}{y_{max} - y_{max}}$$

K-Medoid Clustering

Based on the above four attributes, K-Medoid clustering was initially performed to find a representative data point for every single HCPCS process. This was performed by subsetting the data based on the process and performing an iteration of K-Medoid clustering with only one medoid, chosen to represent each process. While clustering there were three possibilities that could occur with a given subset of data:

1) There is only one data point for a given process, implying that it is the medoid without any additional computational need
2) There is a large number of data points for a given process, to the point where it is computationally infeasible to handle. At this point the data is clustered using random sampling by the *clara* function in R
3) There neither too low nor too high data to process, which is clustered using the *pam* function in R

With the following possibilities in place, medoids were calculated for the states of Illinois, Virginia, California, New York, and Michigan, which were selected randomly. Each state had a different number of processes, ranging from 2458 to 3510. Calculating medoids has several benefits:

1) It provides a clear interpretability into which processes are lacking in discounts and coverage
2) It accurately scales down the data used for clustering – from over 320,000 points from the state of Illinois, approximately 2,800 points were selected. This makes comparing across multiple states more feasible.

All of the medoids that were selected were put into another data frame, which was then used for clustering.

K-Means Clustering for Multiple States

In order to apply K-means clustering, the optimal number of clusters first had to be determined. In order to do so, the Within Cluster Sum of Squares (WCSS) was calculated for dividing the data into anywhere from 2 to 15 clusters:

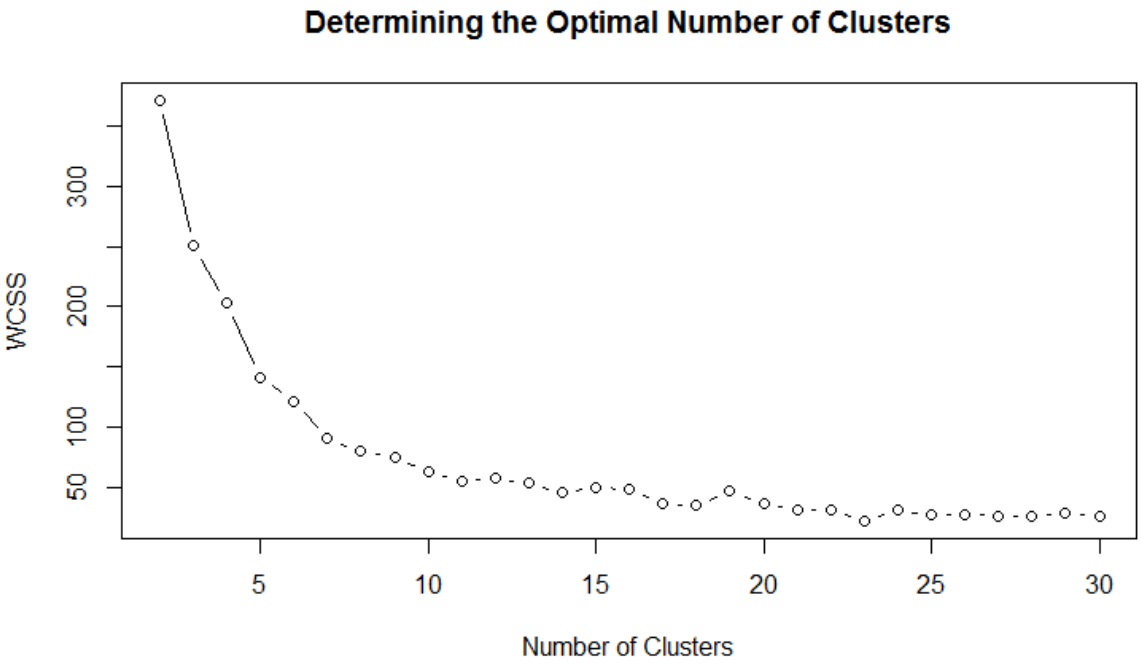
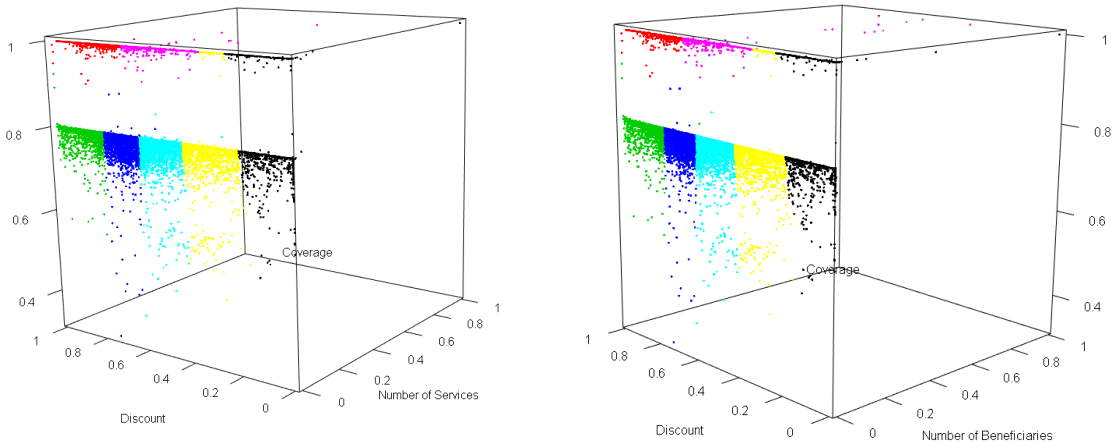


Fig 1: Determining the Optimal Number of Clusters

Based on the graph, the number of clusters selected was 7. Clustering the data points finally yielded the following plots:



Figures 2 & 3: 3-D Cluster Visualizations with Discount, Coverage, and Services or Beneficiaries

The resulting clusters were interesting in several ways – firstly the data seems to be split horizontally into two segments – almost all of the samples had a coverage of 100% or lower than 80%. Furthermore when the coverage is 100%, then the discount doesn't matter as much because the beneficiary is still fully accounted for. However the clusters are divided vertically because of k-means' need to keep clusters of the same size, which isn't the most appropriate choice.

Also, given the lack of variability in the number of beneficiaries and number of services, the processes that seem to be most in need of increased Medicare are those which have lower (<= 0.8) coverage and low (< 0.2) discount. By isolating these processes it is possible to check which ones require attention across all states as well as which ones are unique to a state based on the number of appearances.

Processes Requiring Attention in all States

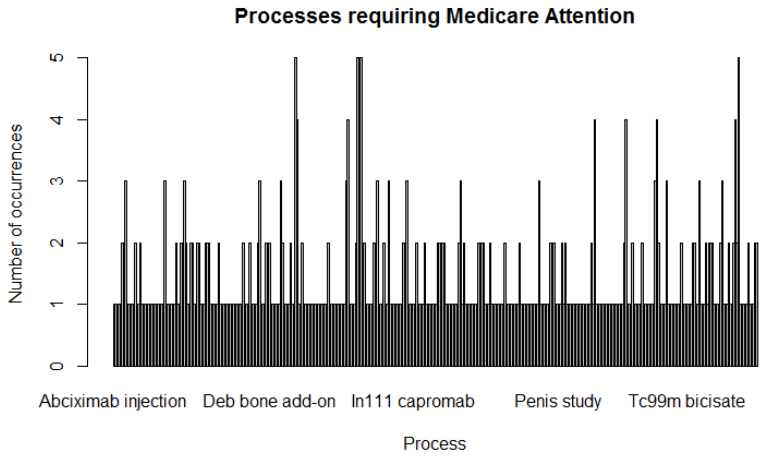


Figure 4: Processes Requiring Medicare Attention

The processes that require Medicare attention across most to all of the sample states (4-5 occurrences) are:

- *Drainage of hematoma/fluid*
- *Eye exam & treatment*
- *Eye exam established part*
- *Eye exam new patient*
- *Trim skin lesions over 4*
- *Drainage of skin abcess*
- *Excision of nail fold toe*
- *Ranibizumab injection*
- *Remove foreign body*
- *Routine footcare pt w lops*
- *Trim skin lesion*

Processes that Require Attention in Specific States

Of the processes that are in the weakest cluster, there are several that also have only one appearance. Processes that appear in this cluster only once but are practiced across all states are also anomalies. Note that for the purposes of this cluster, there isn't a specification that all other points have to be in one cluster. The implicit assumption here is that all other clusters have better Medicare policies for the same process. Based on this, there are several processes that could use improvement, such as:

- *ALS1-emergency in Virginia*
- *Allergy patch tests in Virginia*
- *Anesth dx knee arthroscopy in Illinois*
- *Anesth nerve block/inj in Illinois*
- *Antigen therapy services in California*

The full list will be visible in the console output of the R script.

Conclusion

By first using K-medoid clustering and then using K-means clustering, a sample the US Medicare dataset has been narrowed down to specific processes that need to be improved, based mostly on the discount and coverage rates but also clustered based on the number of beneficiaries and total number of services in the process. There are in total 11 processes that need increased attention across the chosen sample of 5 states and a larger list of processes that need attention in one state but seem to be better covered in others.