Analyzing US Insurance Claim Data: Part II

HW4

Group 5

Roger Hung, Erin Cao, Henry Gao, Stella He, Jiajie Yuan, Robert Malongo

Introduction

In this project, we further explore the insurance claim data and focus more on cohort analysis and cluster analysis. In particular, we use Inpatient and Outpatient data and Revenue code file. We make a narrow study on some particular diagnosis, find the most concentrated clinic chapters among few big players, and do cluster analysis on certain cost categories.

First, we select patients diagnosed as RA (Rheumatoid Arthritis) and count the number of people diagnosed and those not. We later do Fisher Exact test on the gender difference and calculate the inter-quartile range of the costs.

Then, we study on some particular procedures performed by hospitals. We make informed guess as which MDC would be done more generally by most of the hospitals and which one tends to be highly concentrated among specialized high technology medical centers. We later perform analysis to test whether our guess is true.

Lastly, we cluster the cost of payers by the average charge of each DRG. In this study, we filter the admissions to only important DRGs and exclude those rather low dollar value services.

Results and Discussions

Question 1: Study of A Disease Cohort

In general, a study of a disease cohort follows some basic steps including making sure the cohort and sub-cohorts, exploring the cohort's demographics, studying the patterns and other features researchers are interested in. In our case, we studied patients with **Rheumatoid Arthritis (RA)** following the same procedures.

Firstly, we filtered the data by **RA_ICD10_Codes** we already had and found that 976 patients were suffering from common RA and 30 patients were diagnosed having other RA with systemic involvement. Considering that each patient could have more than one type of RA, the total number of diagnoses of common RA and other RA was 981 and 31 respectively. We also found three of the most common RA for each sub-cohort (Table 1.1a & 1.1b). The number shows that most of RA diagnoses are unspecified and the most common complication of RA is rheumatoid lung disease.

Common RA		
ICD-10 Codes	RA Type	Frequency
M069	Rheumatoid arthritis, unspecified	909
M0579	Rheumatoid arthritis with rheumatoid factor of multiple sites without organ or systems involvement	17
M059	Rheumatoid arthritis with rheumatoid factor, unspecified	8

Table 1.1a Top 3 RA type found in common RA sub-cohort

Other RA with systemic involvement		
ICD-10 Codes	RA Type	Frequency
M0510	Rheumatoid lung disease with rheumatoid arthritis of unspecified site	21
M05671	Rheumatoid arthritis of right ankle and foot with involvement of other organs and systems	2

M0519	Rheumatoid lung disease with rheumatoid arthritis of multiple sites	2

Table 1.1b Top 3 RA type found in Other RA sub-cohort

After identifying the cohort, we wanted to test the pattern in RA prevalence between genders. We found in the common RA sub-cohort, it was obvious that there was gender bias in RA prevalence as the RA prevalence in female are much higher than male(Figure 1.1a), but regarding other RA sub-cohort we could not be sure if the gender bias existed (Figure 1.1b). Therefore, we conducted Fisher Exact Test and the results show that the difference in RA prevalence between males and females is statistically significant (Table 1.2a). However, the gender bias of other RA prevalence is not significant due to the test results (Table 1.2b).

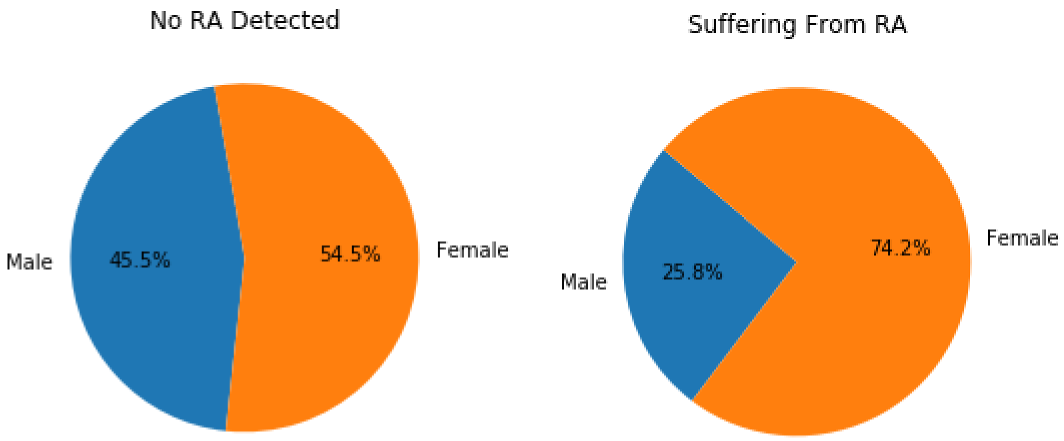


Figure 1.1a common RA prevalence between genders

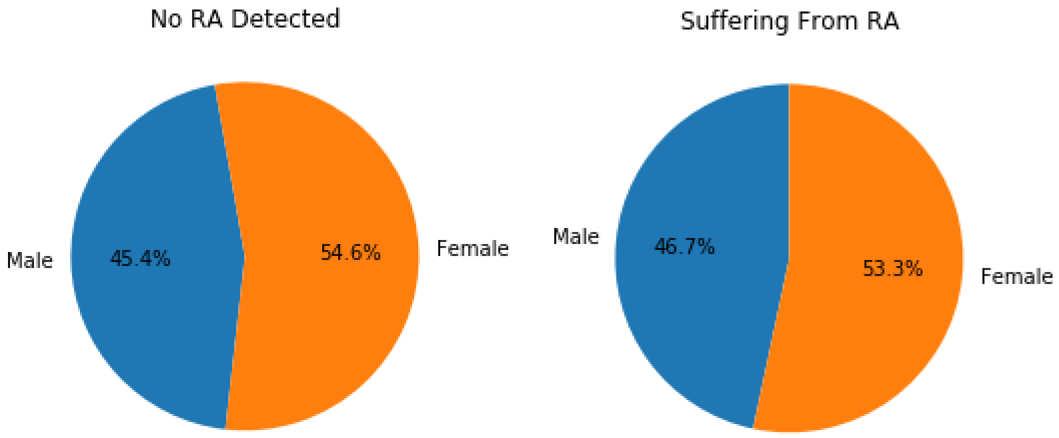


Figure 1.1b other RA prevalence between genders

	Male	**Female**	***Total***
No RA Detected	168150	201496	**369646**
Suffering From RA	252	724	**976**
Odds Ratio	**2.3976**		
P-Value	**1.4458e-36**		

Table 1.2a common RA prevalence between genders with Fisher Exact Test

	Male	**Female**	***Total***
No RA Detected	168388	202204	**370592**
Suffering From RA	14	16	**30**
Odds Ratio	**0.9517**		
P-Value	**1.0**		

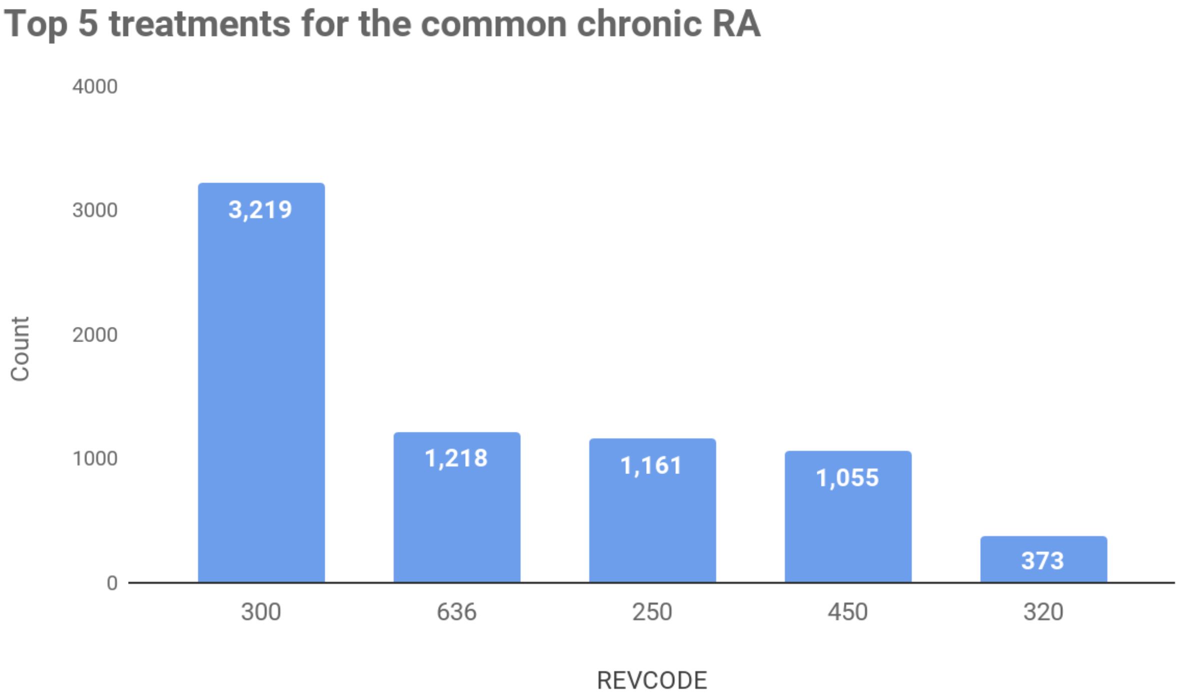
Table 1.2b other RA prevalence between genders with Fisher Exact Test

The interquartile range(Q3-Q1) of the **charges** is calculated to get us a better understanding of the statistical dispersion of the charges variable. Before deriving the interquartile range of charges, we first calculated the first(Q1) and third(Q3) quartiles of the costs and subtracted Q1 from Q3 to get the final result.

	Q1	**Q2**	**Q3**	**Q4**
	25%	50%	75%	100%
	682.48	1521.62	3440.18	227311.78
interquartile range (Q3-Q1)	**2,757.7**			

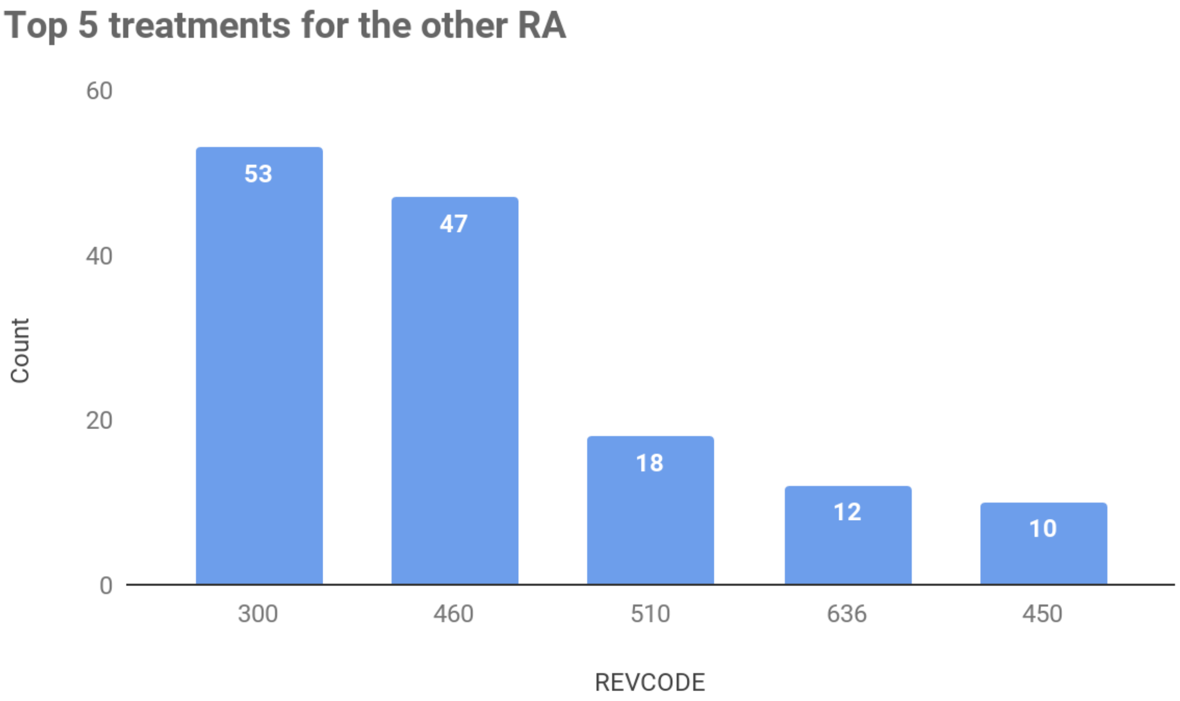
Table 1.3 quartile values of charges

Based on prior analysis, we proceeded to link two sub-cohorts to the Revenue Code file, looking for the top 5 services for treatment for the RA. The result turned out that the most common services provided for treatment for the common chronic RA are Laboratory - Clinical Diagnostic, Pulmonary Function, Clinic, Drugs Require Specific ID: Drugs requiring detail coding, and Emergency Room, in that order. The top 5 common services for treatment for other Rheumatoid Arthritis with systemic involvement are Laboratory - Clinical Diagnostic, Drugs Require Specific ID: Drugs requiring detail coding, Pharmacy, Emergency Room, and Radiology - Diagnostic.



300 : Laboratory - Clinical Diagnostic
636 : Drugs Require Specific ID: Drugs requiring detail coding
250 : Pharmacy
450 : Emergency Room
320 : Radiology - Diagnostic

Figure 1.3a common chronic RA



300 : Laboratory - Clinical Diagnostic
460 : Pulmonary Function
510 : Clinic
636 : Drugs Require Specific ID: Drugs requiring detail coding
450 : Emergency Room

Figure 1.3b other RA

Question 2: Concentration of Major Diagnostic Category MDC

Before doing the analysis, we guessed that MDC 14 (Pregnancy, Childbirth and Puerperium) would be done more generally by most hospitals while MDC 1 (Diseases and Disorders of the Brain and Nervous System) tends to be highly concentrated among specialized high technology medical centers because it requires more complex technology and specialized medical staff.

In order to testify our guess above, we used data file VTINP16upd to perform analysis. First, we counted the total number of patients under MDC 1 and the number of patients in each hospital under MDC 1. Then, we did the same thing with MDC 14. Moreover, we repeated this process by total charge in $. Both methods should give us the same result.

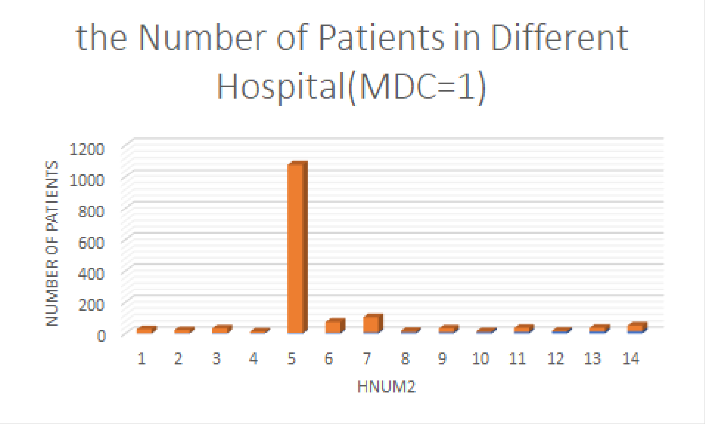


Figure 2.1a Number of Patients with MDC 1

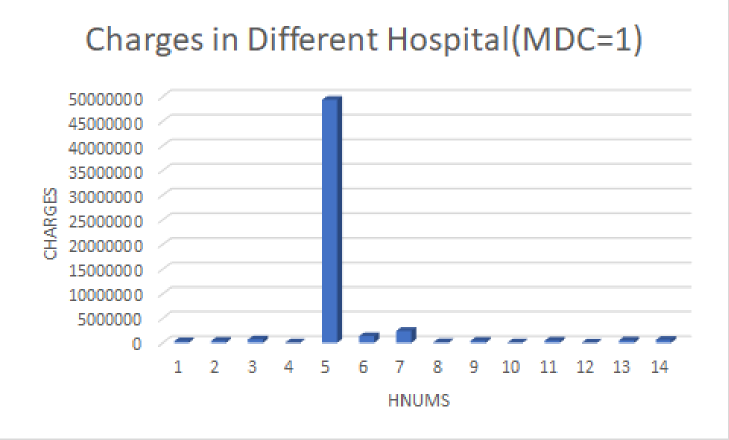


Figure 2.1b Charges of Patients with MDC 1

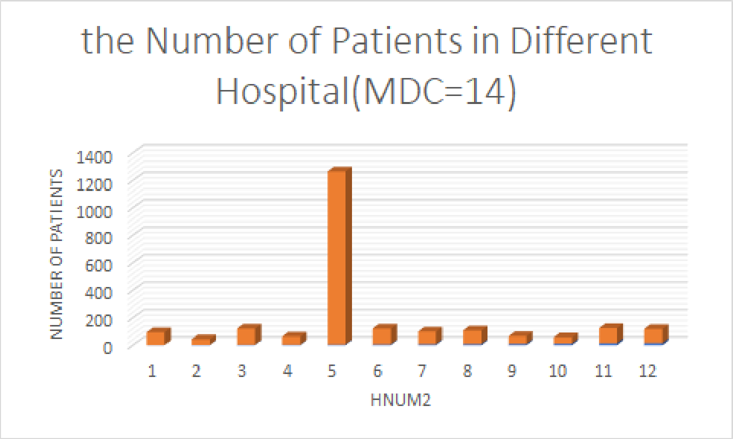


Figure 2.2a Number of Patients with MDC 14

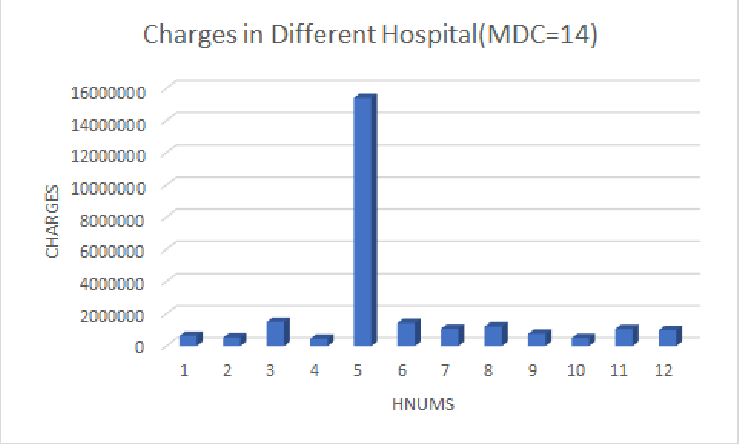


Figure 2.2b Charges of Patients with MDC 14

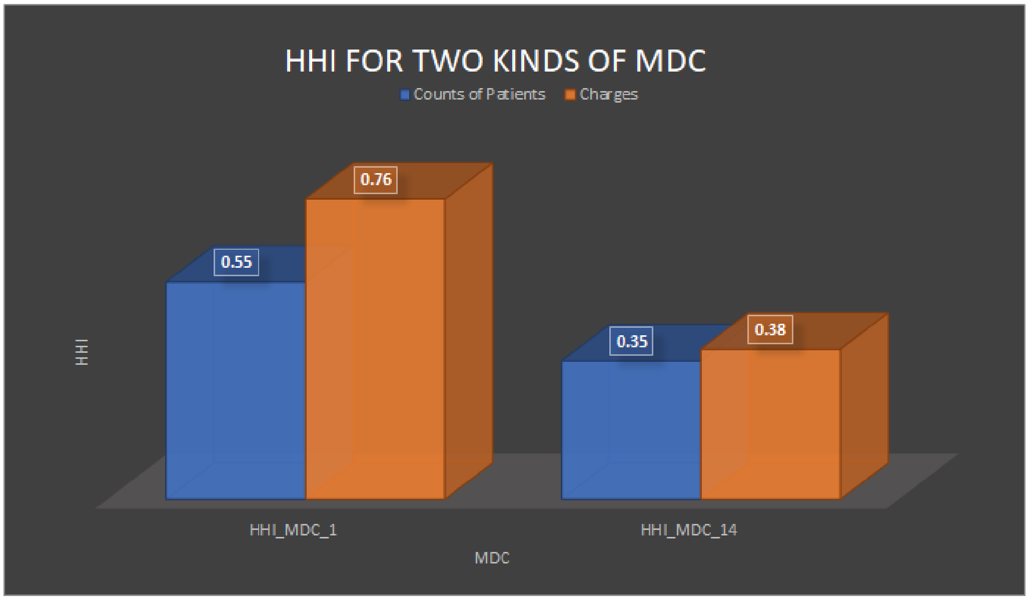


Figure 2.3 Number of Patients and Charges HHI in both MDC

After getting the number of patients and total charges in each hospital under MDC 1 and under MDC 14, we visualized the results into Figure1.1a, Figure1.1b, Figure1.2a and Figure1.2b. However, Hospital number 5 seems to play a very important role in both MDC 1 and MDC 14. In order to better understand the market share, we calculated the HHI for both number of patients and charges in MDC1 and MDC14. As shown in Figure 2.3, the HHI for both number of patients and charges are higher in MDC1 than those in MDC14. As a result, MDC1 is much more concentrated. Hospital number 5, which is the University of Vermont Medical Center, holds the lion's share in the market by taking 73.85% admission counts and 86.90% medical charges alone.

University of Vermont Medical Center is an academic medical center located in Burlington, Vermont. It serves as both a regional referral center and a community hospital. The five specialty areas ranked as High Performing are Orthopedics, Neurology, Neurosurgery, Gynecology and Nephrology. Our initial guess was right: patients with more complex disease such as brain disorder tend to be highly concentrated in the referral hospitals/centers which can take more challenge in medicine.

Question 3: Clustering costs

A diagnosis-related group (DRG) is a patient classification system that standardizes prospective payment to hospitals. The DRG system categorizes hospitalization costs and determine how much to pay for a patient's hospital stay. The DRG assigned to hospitalization depends on the following parameters:

- Principal diagnosis
- Secondary diagnosis(es)
- Surgical procedures performed
- Concurrent illnesses and complications
- Patient's age and sex
- Discharge status

For this question we carried out K-means cluster analysis to classify inpatient DRG admissions using one cost variable .i.e **PCCR_OR_and_Anesth_Costs** (sum of the average **Operating Room and Anesthesiology** costs per patient). Prior to carrying out k-means clustering we prepared out data by following the instructions in the assignment prompts. The major data cleaning step included

- Choosing only DRGs between 20 and 977 from the Inpatient database
- Dropping revenue charges below $100 from the revenue database

- Combining revenue databases with inpatient databases

K-means clustering is a type of unsupervised learning commonly used on data without defined groups. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

- The centroids of the K clusters, which can be used to label new data
- Labels for the data (each data point is assigned to a single cluster)

Each centroid of a cluster is a collection of feature values that define the resulting groups. The graph on the next page shows the results of our K-means cluster analysis:

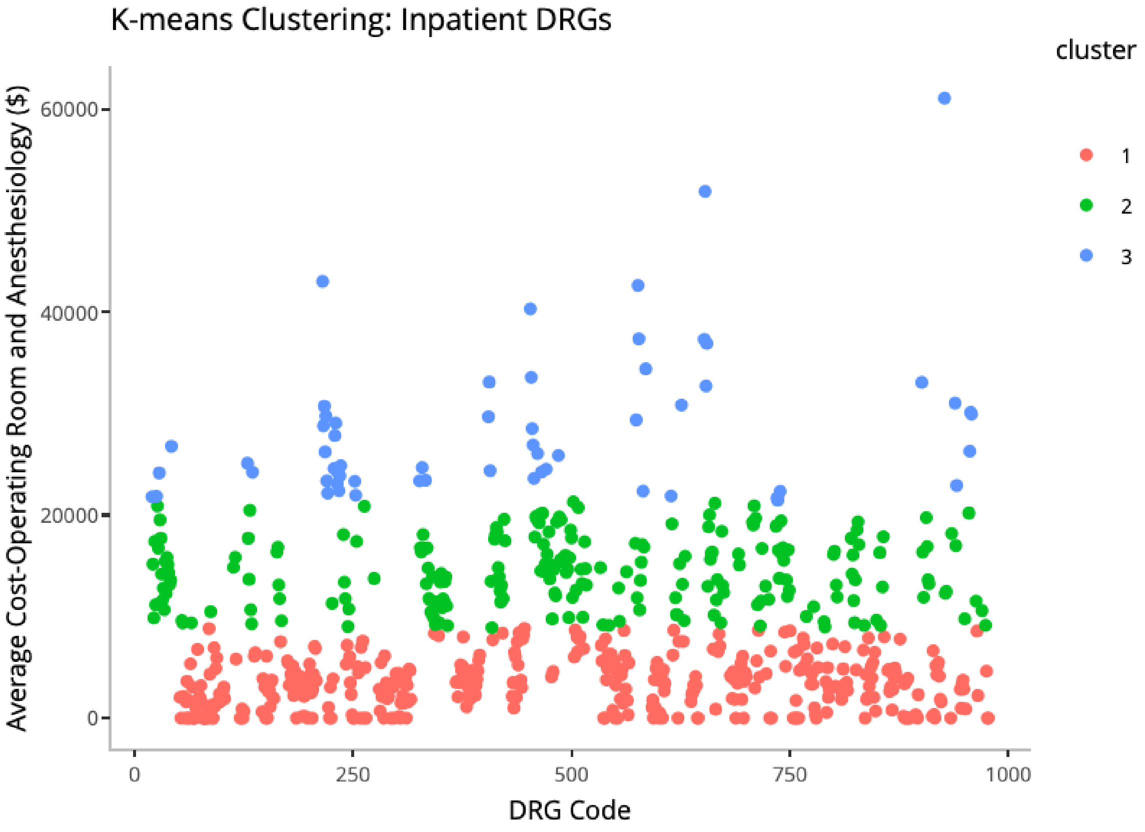


Figure 3.1 K-means Clustering Results

The three clusters above seem to show certain properties of each patient DRGs. In particular, we found evidence that the DRG codes clustered by the frequency of the specific DRGs (and diseases associated with the DRG) which ultimately explains the complexity of the procedure the patient received and the length of stay at the hospital. This is summarized in the flowchart below:

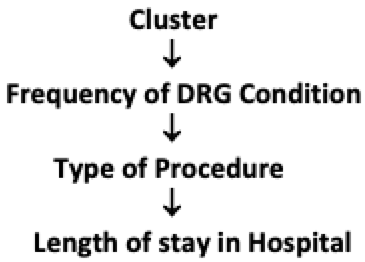


Figure 3.2 Analysis Flow

Note that we obtained the following results: Cluster 1 (408 DRGs), Cluster 2 (234 DRGs), and Cluster 3(61 DRGs). Our analysis shows that Cluster 1 has conditions that are *common* among the general population while Cluster 2 and Cluster 3 have *rare* and *extreme* conditions respectively. The DRGs in cluster 1 include infections, respiratory, reproductive, and rudimentary heart illnesses with concurrent conditions. Diseases associated with these DRGS are fairly common across all age groups in the US. on the contrary, Cluster 3 has complex and rare DRG conditions such as skin grafting, heart implants, and brain procedures.

The chart below is plotted based on the 408 most frequent DRGs based on utilization data provided by CMS and private insurance from more than 3,000 hospitals using the Inpatient Prospective Payment System (IPPS)[1]. Assuming our cluster analysis interpretation is correct, we would expect a higher proportion of these DRGs to be from Cluster 1

[1] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2017

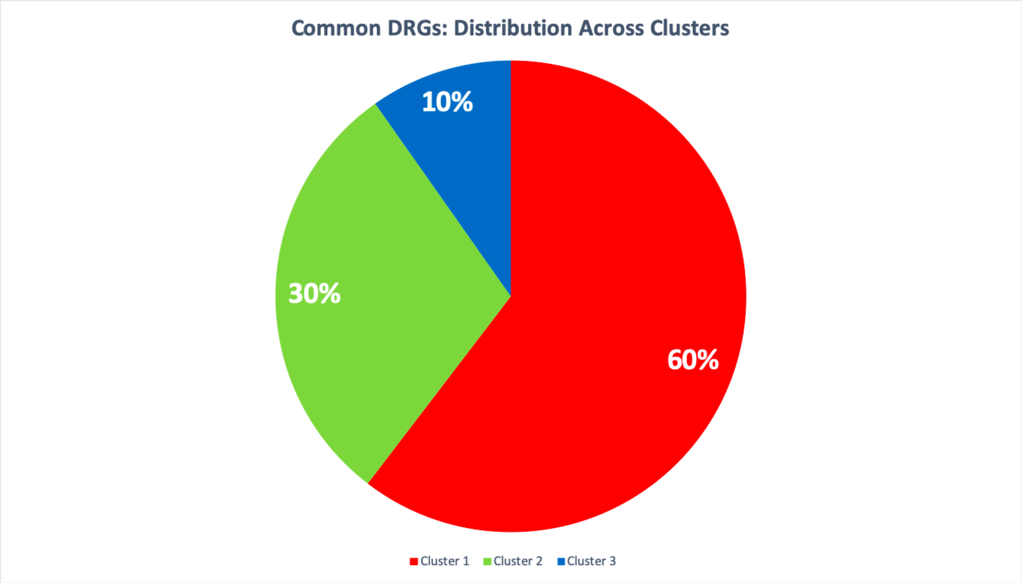
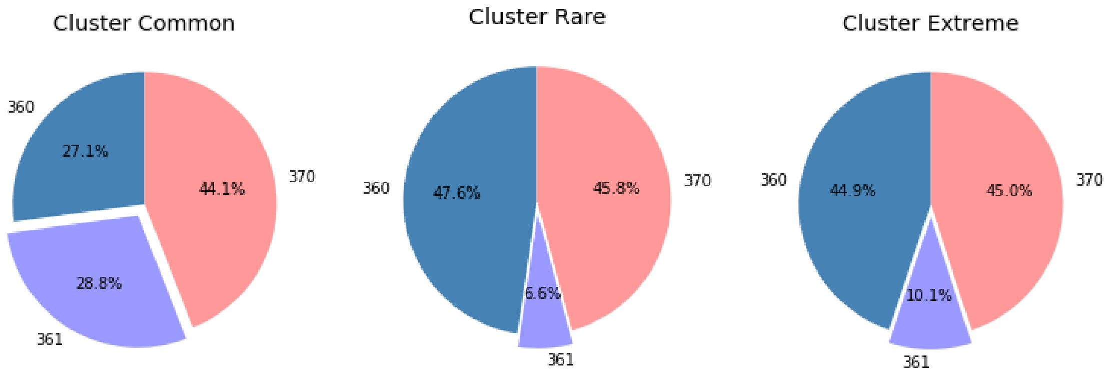


Figure 3.3 Common DRGs: Distribution Across Clusters

60% of the most commonly billed DRGs are categorized in our Cluster 1. DRGs in Cluster 2 and Cluster 3 account for lower shares at 30% and 10% respectively. These results validate our assumption that these clusters group DRGs based on the frequency of diseases associated with them.



360: Operating Room Services
361: Operating Room Services: Minor surgery
370: Anesthesia
Figure 3.4 The portion of service provided (Revcode) among three clusters

However the complexity of procedures carried out on a patient is a function of the frequency of the diagnosis. Rare diagnosis requires more complex procedures resulting in higher costs. The pie charts show the patterns of services provided across the three clusters (Figure 3.4), where the portion of service provided is calculated by counting the number of inpatients per the type of service (Revcode) they accepted in each cluster. From the pie charts, interestingly, we found that a larger part of patients in Cluster Common accepted operating room for only minor surgery than the other clusters, which implicated that the patients in Cluster Common might have less serious or mild diseases and just need minor surgery treatment that costs a small amount of money. Nevertheless, we could not be sure about the difference between Cluster Rare and Cluster Extreme merely on these pie charts.

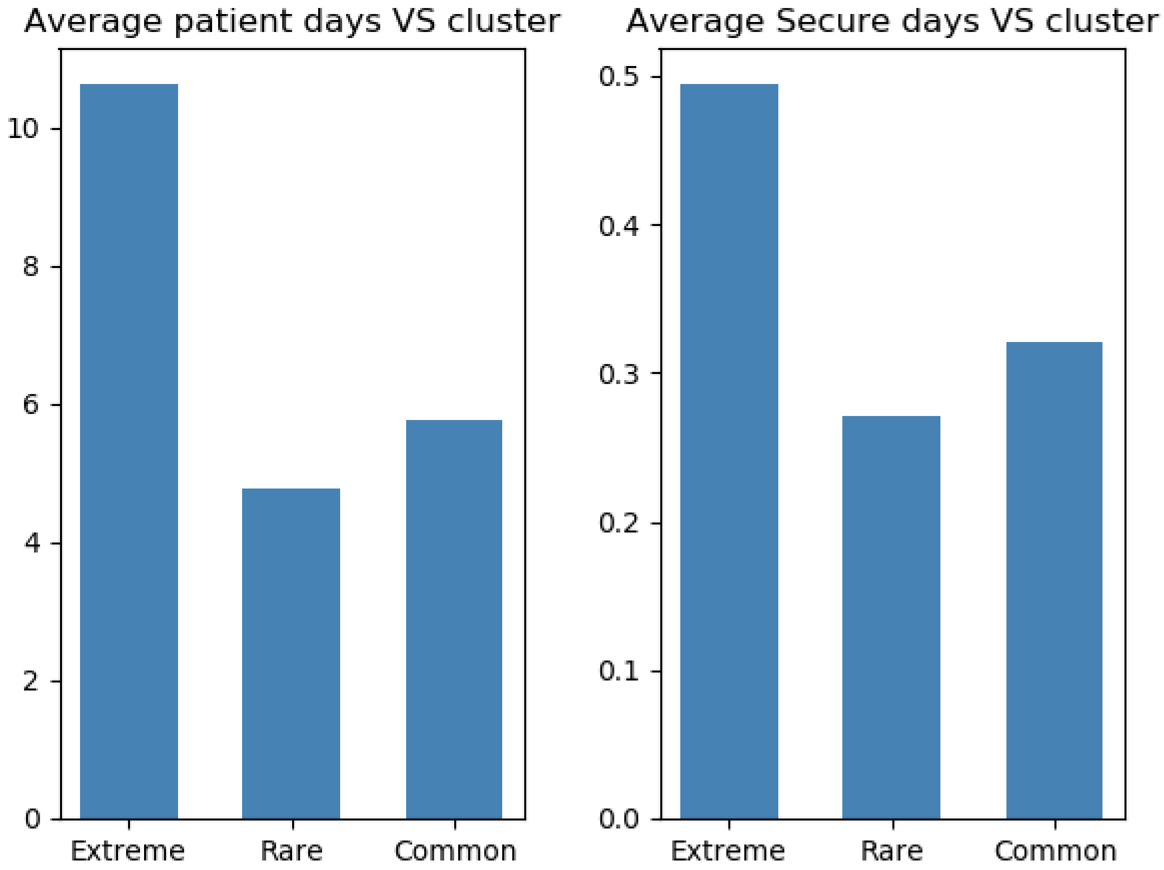


Figure 3.5 Average days patients stay in the hospital and in the special care unit (e.g. ICU)

Accordingly, we dug deeper into inpatients dataset, and at this time, we focused on the days patients stayed in the hospital or in the special units. Fortunately, we calculated the average patient days and found that patients in Cluster Extreme stay much longer in the hospital than Cluster Rare and the average secure days showed a similar pattern (Figure 3.5). Furthermore, we used t-test to validate our findings and got p-value of 1.56e-201 and 0.004 respectively, which means the difference of average days patient stay between Cluster Extreme and Cluster Rare is statistically significant. Therefore, we replenish our story with that the patients in Extreme Cluster may have serious diseases even with some complications due to which they need to stay much longer in the special unit and hospital than patients in Cluster Rare. It would be confusing that patients in Cluster Common stay a little longer than Cluster Rare, but by T-test, we found the difference was not significant so we could ignore that and also we were not hoping to differentiate Cluster Common and Cluster Rare with these charts.

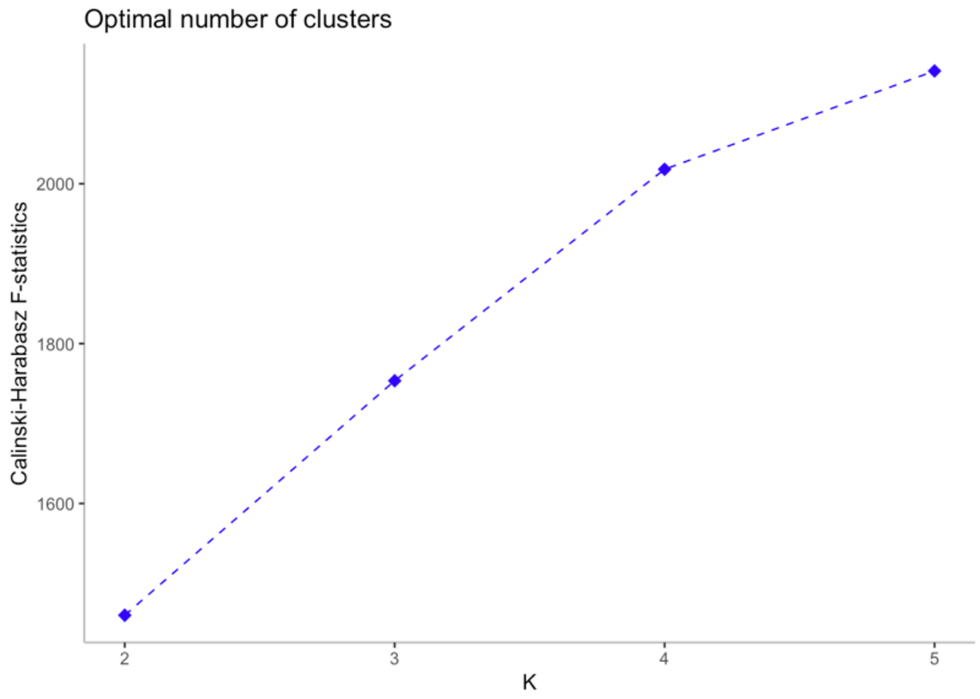


Figure 3.6 F-statistic across clusters (k = 2, 3, 4, 5)

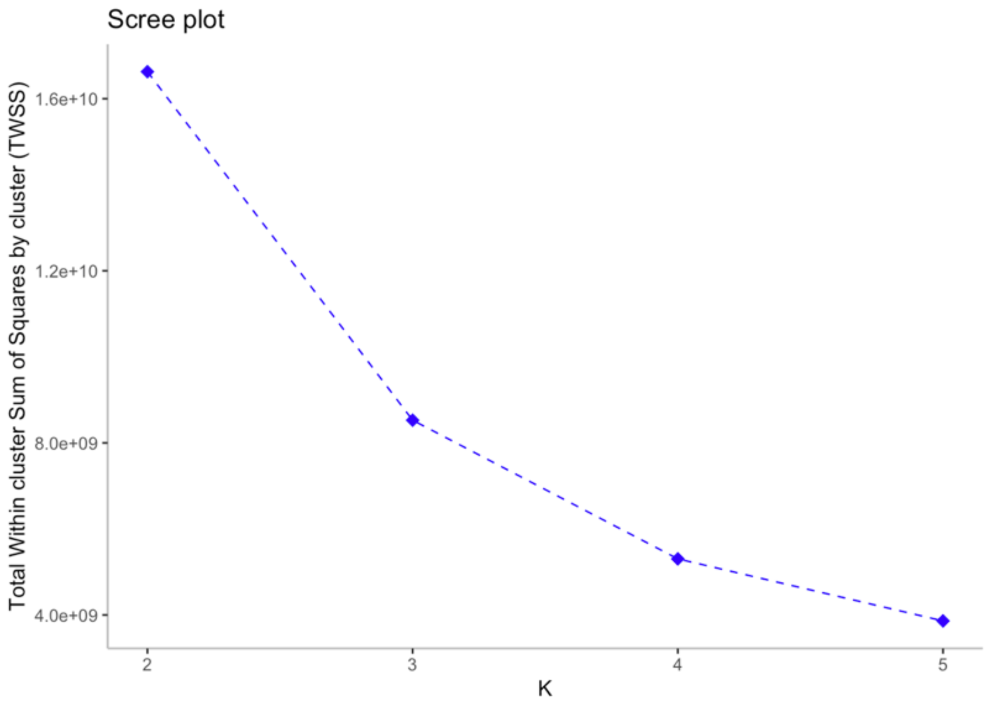


Figure 3.7 Scree plot (k = 2, 3, 4, 5)

Conclusion

Cohort study enables us to conduct analysis on a group of people who share a defining characteristic. We identify the common RA group and other RA with systemic involvement group and learn from fisher exact test result that the gender difference in common RA prevalence is significant but not significant among other RA groups. We also conduct ltwo sub-cohorts and find the top 5 services for treatment for the RA and other RA respectively. The two groups share the top one treatment, Laboratory - Clinical Diagnostic, and have different other most seen treatments.

We study the concentration of two Major Diagnostic Category, MDC 1 (Diseases and Disorders of the Brain and Nervous System) and MDC 14 (Pregnancy, Childbirth and Puerperium), and find that MDC 1 has a particular concentration on specialized high technology medical centers and MDC 14 is matched more with general hospitals.

Based on the average cost of Operating room and Anesthesiology, we generate 3 clusters, representing Common, Rare and Extreme respectively. Common group DRG has the most number of DRGs and number of patients and Extreme group has the least. In terms of the secure days and patient stay, patients with DRG classified as Extreme tend to stay in hospital the longest time which is consistent with our conclusion.

Recommendation

According to our analysis, some diseases such as disorder in brain and nervous system that require more complex technology and more experienced professionals would tend to be treated highly concentrated in certain hospitals. In this case, we recommend insurance companies and hospitals to be aware of these kind of “high capability” hospitals and make them more accessible to patients who’s in complex condition.

Moreover, we recommend insurance companies to group patients into different groups by the severeness of their condition and set up different plans accordingly. When conducting background checks, insurance companies should pay more attention on heart disease since it plays an important role in extreme conditions.

Reference

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2017
https://hmsa.com/portal/provider/zav_pel.fh.DIA.650.htm

Appendix

- Python code for Q1 and Q3
- Python code for Q2

END