NOVEMBER 29, 2019

INSURANCE CLAIMS DATA & ANALYTICS PART II

HEALTHCARE DATA MINING AND DATA ANALYSIS

QINGYUE SU
MSBA, BRANDEIS UNIVERSITY
qingyuesu@brandeis.edu

Table of Contents

Insurance Claims Data & Analytics Part II

Introduction

This assignment continues concentrating on the analytics on the most important database in market for health, the Insurance Claim Data also known as Administrative Claim Data. These databases consist of inpatient discharge data, outpatient procedures and services data, and emergency department data. I use the available database from year 2016.
I firstly study the patients with Rheumatoid Arthritis (RA), identify its cohort and major sub-cohorts; explore the cohort's demographics and gender differences; calculate the costs and study the service utilization.

Then I turn to the medical center part, make assumptions and calculate the HHI index to investigate the concentration of the care for the two MDCs (MDC1, MDC14), and examine the lion share hospital. Finally, I conduct a cluster analysis of the hospital DRGs using cost categories. I focus on studying and exploring the clustered DRGs and make interpretation.

After all these works, I get familiar with the Insurance Claim Data and are ready for deeper analysis.

Study of A Disease Cohort

In this part, I will go deep into the discussion of RA. First, I will first analyze the common types of RA. Next, I will analyze the demographics of RA. Last, I will focus on the related costs and treatments of RA.

Step 1: Identify the RA cohort using the outpatient file.

In this section, I mainly divide RA into two types, one that only affects the joints, and the other that is relatively severe and affects the vascular and visceral system. At first, I select the patients who are diagnosed with at least one category of RA of each type. For future use, I also include demographic information about the patients in the column.

	hnum2	sex	Uniq
1	5	2	3469
2	5	2	3747
3	5	2	14845
4	5	2	17180
5	5	1	17860
6	5	1	21606
7	5	2	24499
8	5	2	26942
9	5	2	31326
10	5	1	41545
11	5	1	41671
12	5	2	43660
13	5	1	46928
14	5	2	47358
15	5	2	48638
16	5	2	48936
17	5	1	50425
18	5	1	51051
19	5	2	55037
20	5	2	55421
21	6	2	57678
22	6	1	61965
23	6	2	62050
24	6	1	65577

Showing 1 to 24 of 976 entries, 3 total columns

table1: The RA cohort

	hnum2	sex	Uniq
1	5	1	89770
2	5	2	101488
3	5	2	127067
4	5	2	172527
5	5	2	173928
6	16	1	256318
7	5	1	468716
8	5	1	617255
9	5	1	694041
10	5	2	742535
11	12	1	754146
12	8	2	798736
13	16	1	910288
14	16	2	921450
15	14	1	931784
16	4	2	939269
17	4	1	948290
18	5	1	984326
19	5	2	1103798
20	5	2	1215473
21	5	2	1218171
22	5	1	1218336
23	5	1	1218352
24	8	2	1628990

Showing 1 to 24 of 30 entries, 3 total columns

table2: The other RA cohort

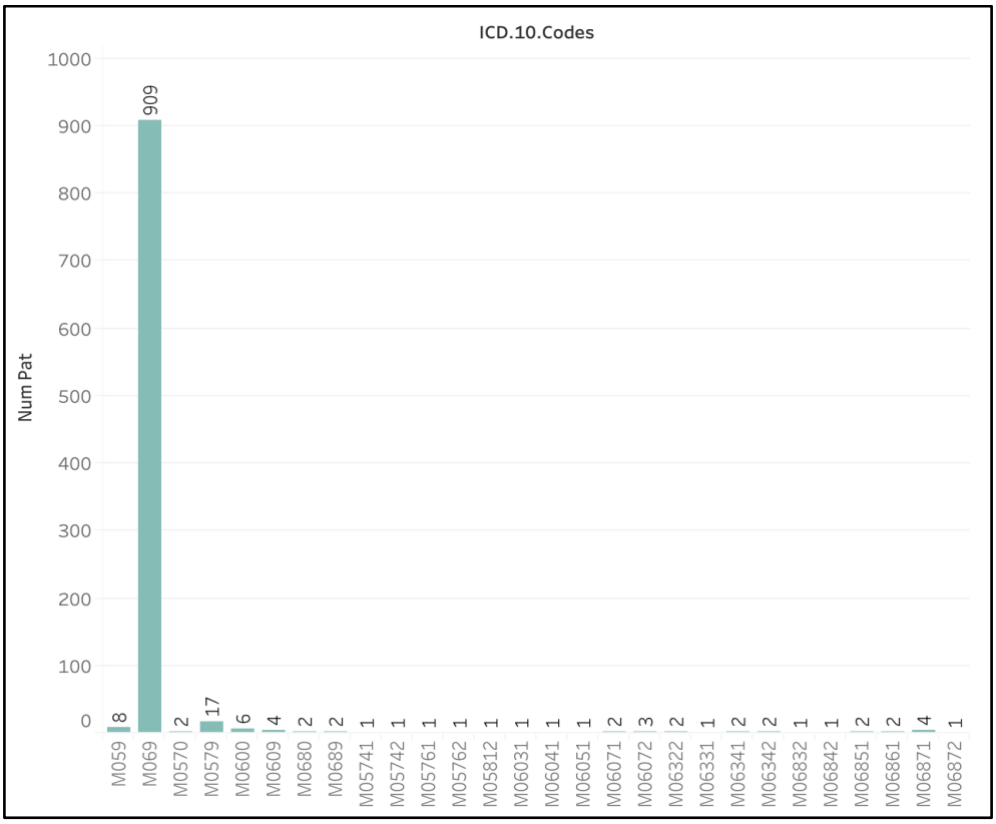
Step 2: Identify the most common types of RA

In this step, I will identify the most common categories of RA of two types.
Among RA types that only affect joints, I can tell from the graph, the most common categories of RA are M069, M0579, M059.

- M069: Rheumatoid arthritis, unspecified
- M0579: Rheu arthritis w rheu factor mult site w/o org/sys involved
- M059: Rheumatoid arthritis with rheumatoid factor, unspecified

Rheumatoid arthritis is a hereditary systemic disease that usually does not affect only a single side of the joint, but rather affects the bilateral side of the joint. In addition, because rheumatoid arthritis is a systemic disease, it may affect multiple joints and cause deformities. Therefore, usually, when patients have symptoms and come to the hospital for diagnosis, they will be classified according to the affected area.

M069 will be a more general ICD for RA, so the number of patients under this classification will be larger. Patients are usually classified as M069 if they are suspected or have mild symptoms of RA. When the patient's symptoms are severe and have affected multiple joints, the patient will be classified as M0579. However, RA is not a common disease, and multi-joint RA is especially rare. Therefore, there will be a significant difference in quantity from M069. As mentioned above, RA is a hereditary disease, it can be known whether the patient has the genetic factor of RA in the detection of RF and anti-CCP antibodies. Blood tests can also be used to determine if a patient is a high-risk group for RA. But patients will not go to the hospital to check whether they are genetically carriers of RA until the symptoms are obvious and affect patients repeatedly. Therefore, patients who will go to the hospital for risk assessment are usually those with suspected themselves RA or those with RA patients at home. This group does not necessarily have symptoms of RA, but they need to return to the hospital regularly for follow-up.

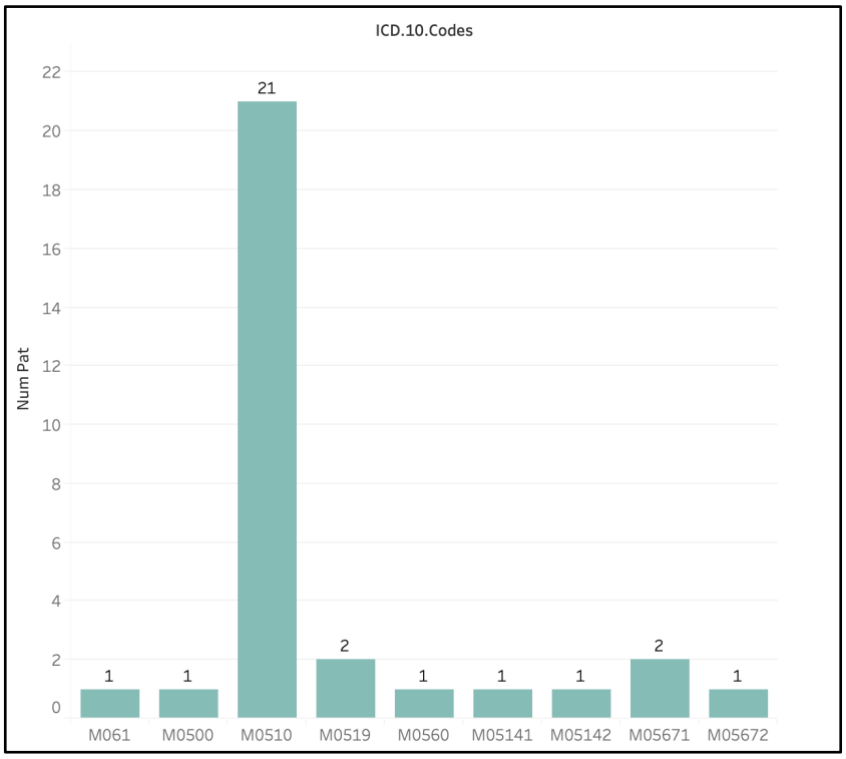


graph1: Distribution of RA patients among different MDCs

Among the more severe types of RA that affect the visceral and vascular systems, I can tell from the graph that the most common types of RA are M0510, M0519, M05671.

- M0510: Rheumatoid lung disease w rheumatoid arthritis of unsp site
- M0519: Rheu lung disease w rheumatoid arthritis of unsp shoulder
- M05671: Rheu arthrit of right ank/ft w involv of organs and systems

RA patients with RA that affect the visceral and vascular systems are very rare, and they usually combine joint problems. When RA affects the internal organs, it may cause inflammation and deformity, affecting overall body function and may even cause death.



graph2: Distribution of other RA patients among different MDCs

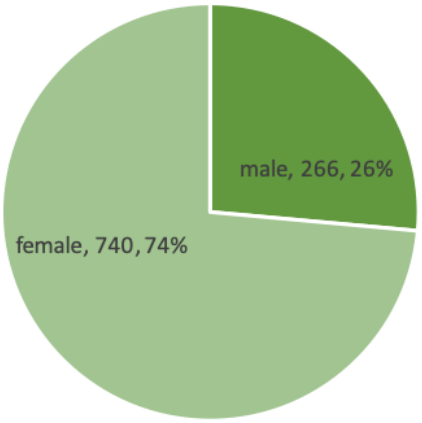
Among the types of RA that affect the viscera, the organs that are most commonly affected are the lungs. RA is often accompanied by interstitial lung disease. As the name of the disease implies, it produces lesions in the interstitial lung. This problem usually does not have symptoms at first, but once it is discovered, the lungs may already be seriously affected. If the patients do not follow up regularly, respiratory failure and death due to pulmonary fibrosis can happen. According to data, the number of patients who have RA in lung and shoulder joints is slightly more than in other categories. In addition, there are also slightly more patients whose right lower extremity and visceral system are affected by RA. However, because patients with visceral systems infected by RA are very rare, the analysis value of this data is not high.

Step 3: Gender differences in RA

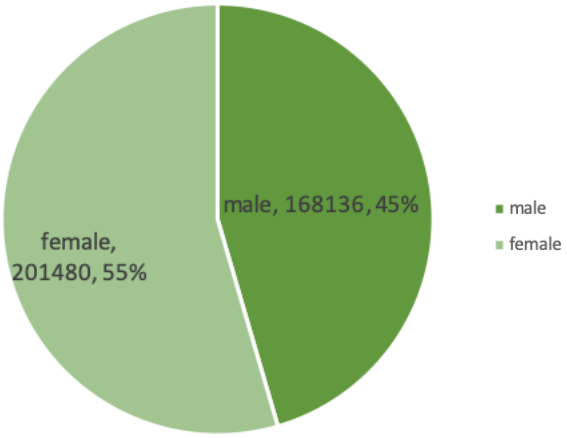
Women are more likely to develop arthritis than men, and RA is no exception. RA is commonly seen in women aged 30~50. From the figure below, I can tell that the ratio of women is greater than that of men. According to research, the ratio of women and men diagnosed with RA is around 3:1. The table below verified the results of the research.

sex_type <chr>	NO RA & Other RA <dbl>	RA & Other RA <dbl>
female	201480	740
male	168136	266

table3: Tabulation of gender difference in RA



graph3: RA & Other RA



graph4: NO RA & Other RA

I assume that women and men are equally likely to develop RA. As a result, I conduct Fisher's Exact Test to verify my assumption. As the result shown below, the p-value under 0.05; therefore, I should reject my assumption. The test shows that women are statistically significantly more common in developing RA than men. Also, the odd ratio is under 1, which also means that women are more vulnerable than men to be diagnosed with RA. There are still no studies to confirm the pathogenesis of RA; as a result, I still can not figure out the reason why women are more vulnerable than men to develop RA.

```
        Fisher's Exact Test for Count Data

data:  t
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3719619 0.4946008
sample estimates:
odds ratio
 0.4294675
```

table4: Result of Fisher's Exact Test

Step 4: Calculate the interquartile range of the costs

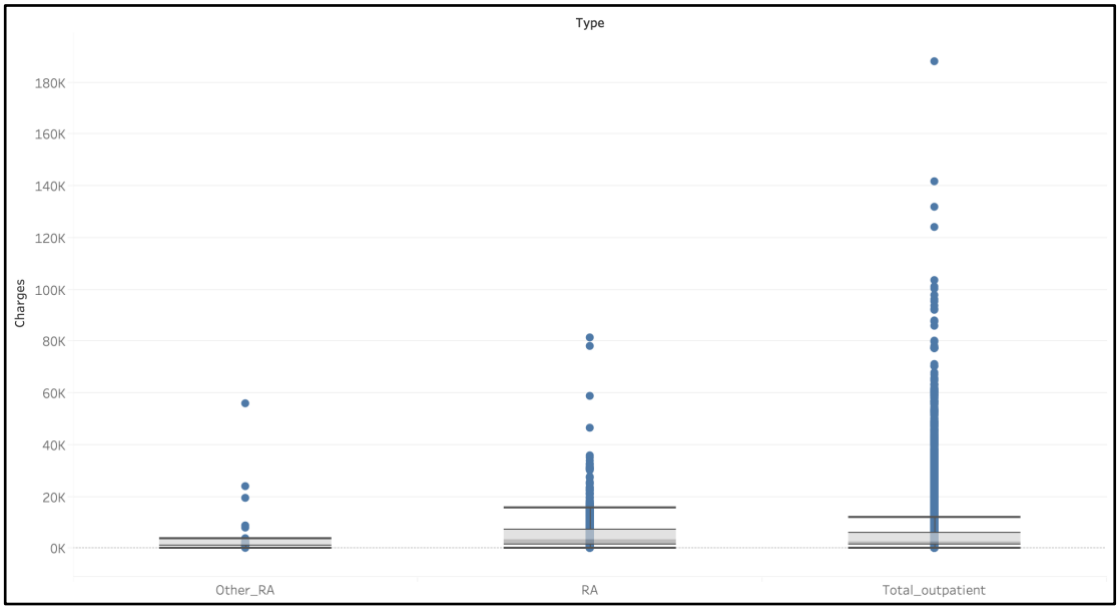
In this step, I will analyze the charge between two types of RA.

I can tell from the graph and table below, the 1st quartile and mean charge of each type is similar. However, the 3rd quartile of charge will be considerably different. The main reason behind this phenomenon is that RA patients with RA that affect the visceral and vascular systems are usually very severe. As a result, the average life expectancy of these patients is also relatively short. Moreover, RA is an immune disease that cannot be cured, its symptoms can only be reduced by long-term medication or conservative treatment. Therefore, if the relative life is shorter, many medical and care costs will be reduced. Another reason behind this phenomenon is that the patients with mild RA may under repeatedly orthopedic surgery to reduce symptoms. Repeatedly surgery can cause a great amount of money. As a result, this is the reason why the 3rd quartile cost of RA patients that only affected joints can be so different from the other type.

Compared with the total cost of other outpatients, the costs of RA patients are relatively high. As I mentioned above, RA is an immune disease that cannot be cured, it is a life-long disease. As a result, the costs for the medication, surgeries, conservative treatment will be relatively higher than other diseases among outpatients.

quantile <chr>	RA_c <chr>	Other_RA_c <chr>	Total_outp_c <chr>
percentile	RA	Other RA	Total outpatient
25%	1234.808	1043.365	682.48
50%	3205.335	1436.19	1521.62
75%	6982.265	2748.97	3440.18
IQR	5747.457	1705.605	2757.7

table5: Quantile of charges



graph5: Distribution of charges among different types of RAs

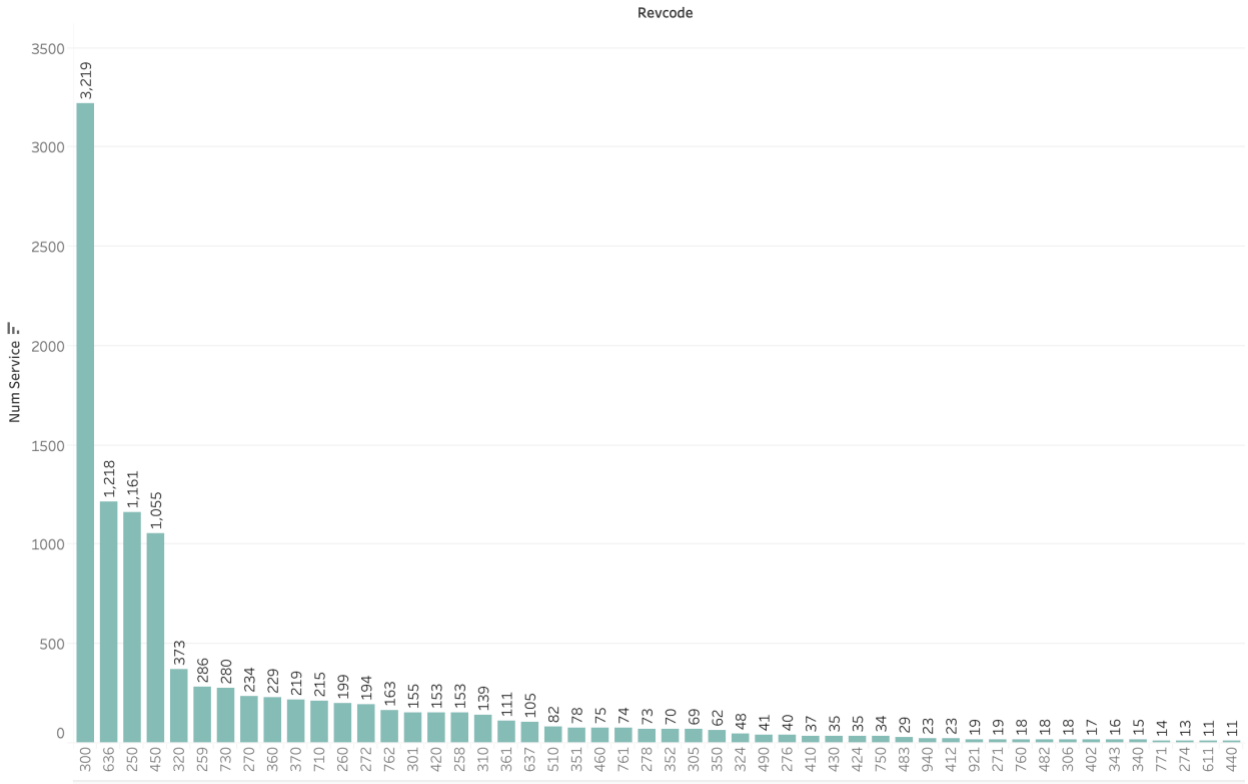
Step 5: Study of service utilization

In this step, I will analyze the top-5 services that RA patients utilize the most.
The top-5 service utilization of RA patients that only have joints problems is 300, 636, 250,450, and 320.

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

The result above is as expected, diseases of the immune system require long-term bacterial culture and special medication. Therefore, a great amount of money will be spent on medication. Also, because RA is an immune disease, there may be some sudden conditions; it is not surprising to enter and exit the emergency room often when the disease occurs. Furthermore, joint diseases require an x-ray or MRI for examination and tracking, so radiology treatment is usually required.

From the above results, I can know that the most common treatments for RA patients are drug therapy and radiological examination and tracking. Also, these patients often have emergency problems and need to be sent to the emergency department.



graph6: Distribution of service utilization of RA (*just shows the top 50 common services)

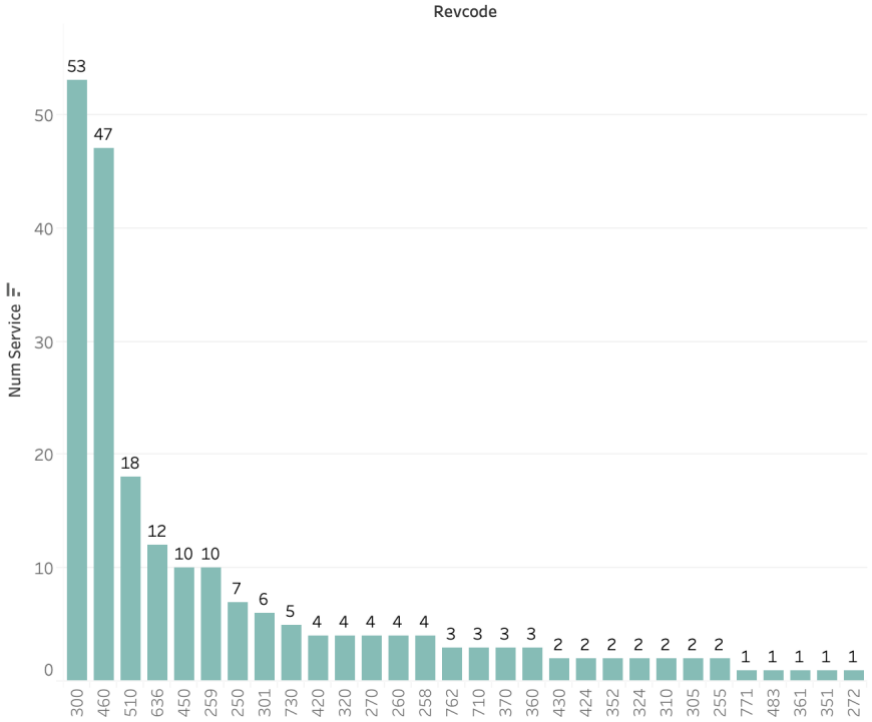
The top-5 service utilization of RA patients that are affected by visceral problems is 300, 460, 510,636, 259, and 450.

- 300: Laboratory - Clinical Diagnostic
- 460: Pulmonary Function
- 510: Clinic

- 636: Drugs Require Specific ID: Drugs requiring detailed coding
- 259: Pharmacy: Other
- 450: Emergency Room

As I mentioned before, this type of patient is more severe than normal RA patients. The internal systems of these patients are likely to be inflamed and deformed, affecting the normal functioning of their bodies. Moreover, as I mentioned earlier, this type of patient is often accompanied by interstitial lung disease. Therefore, in addition to the special medications that normal RA patients need, they also need some assistance and treatment on the cardiopulmonary system.

From the above results, I can know that the most common treatments for this type of RA patients are drug therapy and pulmonary therapy. Also, these patients often have emergency problems and need to be sent to the emergency department.



graph7: Distribution of service utilization of Other RA

MDC Concentration

In this part, I analyze the more concentrated clinical chapter as defined by the Major Diagnostic Category (MDC) in inpatient care among a few big hospitals. I use the inpatient data, compare two MDC: MDC1 Diseases and Disorders of the Brain and Nervous System and MDC14 Pregnancy, Childbirth and the Puerperium.

My guess for the comparison is that MDC 14 would be done more generally by most of the hospitals and MDC1 tends to be highly concentrated among specialized high technology medical centers. The reason is that pregnancy and childbirth will be a common need among patients so most hospitals should be equipped with the related technology and medical staff. On the contrary, brain and nervous system issues are more specific and require high-level abilities, in this case, only several major hospitals are able to deal with such kind of diseases. So MDC1 will be highly concentrated among several medical centers.

Then I calculated the HHI index by patient counts and by total charges in $. From the result, I can confirm my guess that MDC1 is more concentrated while MDC 14 is generally distributed with the index number equals to around 20%.

HHI index:

HHI index	MDC1	MDC14
HHI measured by patient counts	40.71%	21.41%
HHI measured by total charges in $	63.68%	24.54%

table6: HHI index of MDC1 and of MDC14

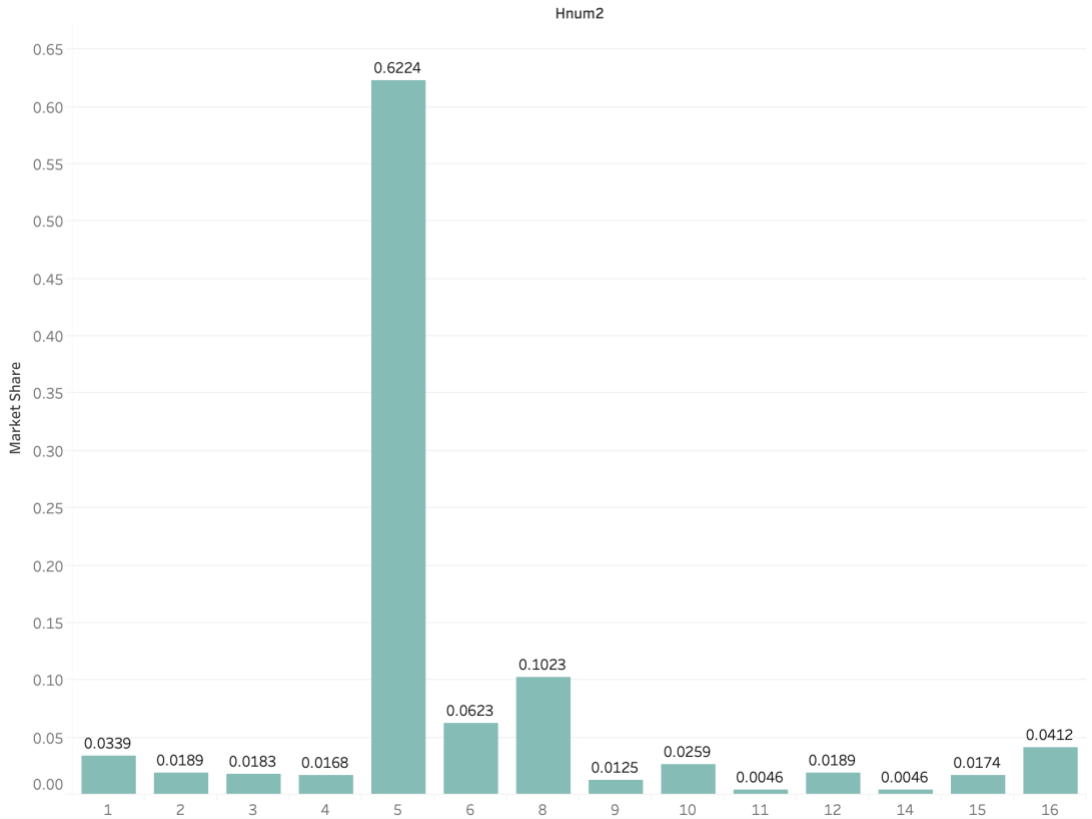
For the more concentrated MDC1, I further analyze the detailed shares among different hospitals.

From the distribution graphs, I can figure out that Hospital 5: University of Vermont Medical Center has the lion share of MDC1 both by patient counts and by total charges in $. For patient counts, its share is 62.24% and for total charges, its share is 79.27%. In this case, the University of Vermont Medical Center plays a major role in dealing with the Diseases and Disorders of the Brain and Nervous System.

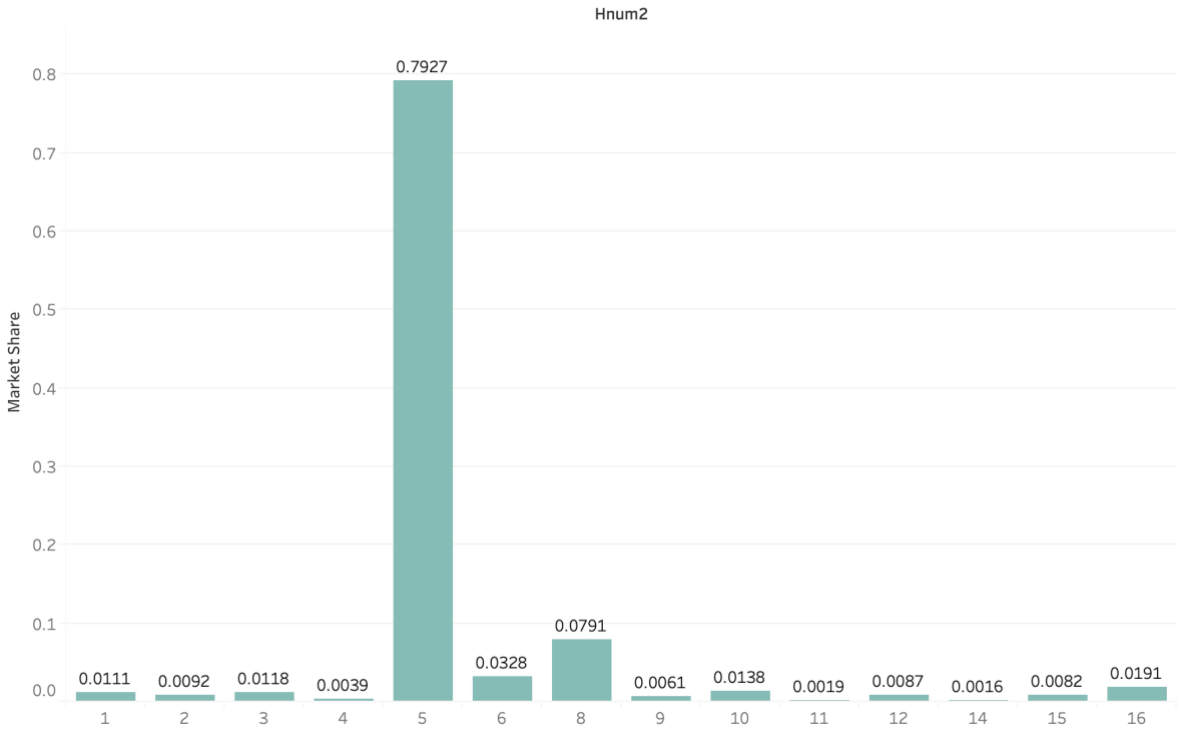
When I looked at more information on the lion share medical center, I noticed that **the University of Vermont Medical Center** is an academic medical center located in Burlington, Vermont, United States. It serves as both a regional referral center (providing advanced care to approximately one million people in Vermont and northern New York) and a community hospital (for approximately 160,000 residents in the Chittenden and Grand Isle Vermont counties).

The medical center has comprehensive surgical services (neurological, cardiac, pediatric) and imaging equipment. It offers leading-edge radiology technology including two Philips Ingenia 1.5T, a Philips Ingenia 3T MRI, a General Electric Signa LX 1.5 tesla system, and a Philips Brilliance 256-slice CT scanner that can produce highly detailed 3D images of the heart, the brain, and tiny blood vessels.

With the information, I know that the University of Vermont Medical Center belongs to “referral hospitals/centers” that can take on any challenge in medicine and operate on virtually any patient no matter how complex the case is. It takes the major market especially in Vermont and northern New York.



graph8: Distribution of market share measured by patient counts



graph9: Distribution of market share measured by total charges in $

After analyzing the results, I confirm my initial informed guess that MDC 14 would be done more generally by most of the hospitals and MDC1 tends to be highly concentrated among specialized high technology medical centers.

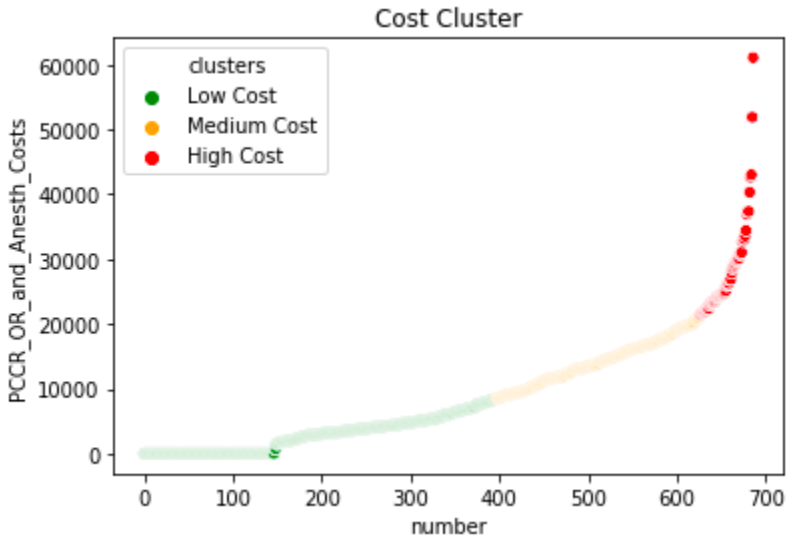
Clustering Costs

In this section, I am going to do a cluster analysis of the cost per diagnosis-related group (DRG). A DRG payment covers all charges associated with an inpatient stay from the time of admission to discharge. The DRG includes any services performed by an outside provider. The dataset is manipulated based on only two revenue centers: operating room and Anesthesiology. Since the category of Anesthesiology is strongly related to any operation, or I could say this is a service attached to operations. Before adding the costs together, I should first substitute the missing values with zeros so that I could get the correct result for the DRGs that have only one of these two cost types. The cluster analysis is aiming at grouping DRGs in such a way that the objects in the same cluster are more similar to each other than to those DRGs in other groups. The clusters are grouped automatically by the k-mean of prices (PCCR) that divided all PCCR costs into

groups. I will explore the reasons and backgrounds of the result of clustering, which are highly related to the real-world healthcare industry. In addition, the process of exploring the similarities and differences among groups allows us to have a better understanding of healthcare insurance claim systems.

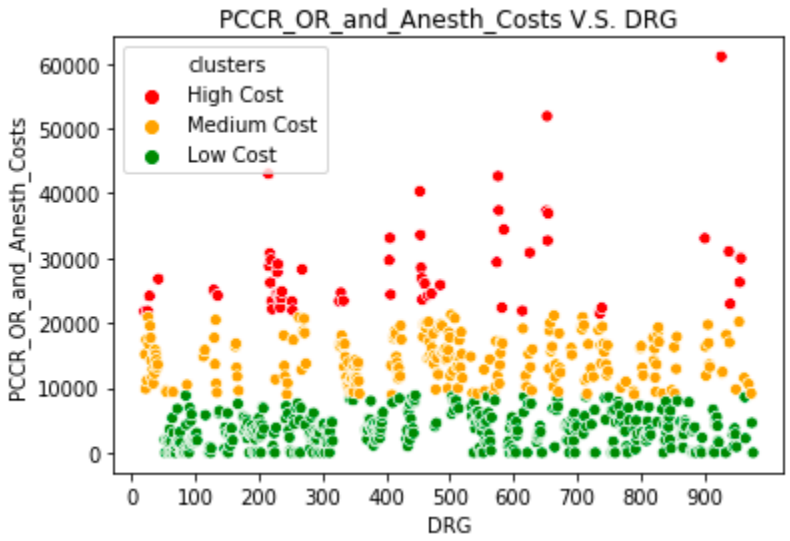
After cleaning data in R, I use Python to cluster average cost data. I try to cluster data into 2,3,4 and 5 clusters and check the F-stat of each cluster. Based on results, F-stat for 2 clusters is 1432.8, F- stat of 3 clusters is 1700.1, F-stat of 3 clusters is 1934.8, and F-stat of 4 clusters is 2144.5. That makes sense. More clusters could make error term smaller, as well as make the model more accurate.

Then I got three clusters of the average cost of Operating Room and Anesthesiology as the figure shows below. Among nearly 700 DIAGNOSIS RELATED GROUPs(DRG), there are about 400 DRGs' average cost of Operating Room and Anesthesiology under 10000, which I could call low cost. 250 DRGs' average cost of Operating Room and Anesthesiology is between 10000 and 22000, we'd like to call it medium cost. Only about 50 DRGs' average cost of Operating Room and Anesthesiology is higher than 22000, which is classified as a high cost.



graph 10: Cost Cluster

For a better understanding of my cluster result, I make some changes to my plot. As shown below, I replace x axial with DRG id to focus on what the cluster means for each DRG id.



graph 11: Cost Cluster of different DRG

In my dataset, there are hundreds of different DRGs, which make it difficult for us to analyze. However, I found DRGs could be divided into several categories and each category has a unique MAJOR DIAGNOSTIC CATEGORY(MDC). So I sort out the range of DRG for each category of MDCs and MDC names and put them together to compare with the cost-DRG figure.

DRG	MDC	MDC_CAT_NAME
20-103	1	BRAIN AND CNS
113-125	2	EYE
129-159	3	EAR, NOSE & THROAT
163-208	4	RESPIRATORY
215-316	5	HEART & CIRCULATORY
326-395	6	DIGESTIVE
405-446	7	LIVER & PANCREAS
453-566	8	MUSCULOSKELETAL
573-607	9	SKIN AND BREAST
614-645	10	ENDOCRINE
652-700	11	KIDNEY & URINARY
707-730	12	MALE REPRODUCTIVE
734-761	13	FEMALE REPRODUCTIVE
765-782	14	PREGNANCY, CHILDBIRTH AND THE PUERPERIUM
789-795	15	NEONATAL
799-816	16	SPLEEN & BLOOD
820-849	17	LYMPHATIC
853-872	18	INFECTION
876-887	19	MENTAL ILLNESS
894-897	20	SUBSTANCE ABUSE
901-923	21	INJURY, TOXIC EFFECTS
927-935	22	BURNS
939-951	23	ALL OTHER
955-965	24	TRAUMA
969-977	25	HIV

Here I am looking into the three clusters separately and make an analysis within each group.

In the high-cost cluster, the DRGs are sorted into 25 MDC categories, then I filtered high-cost MDC types as well as finding common points among them. I find that most of the high-cost surgeries are precision surgeries or surgeries that require implanted devices. For example, CABG or liver and pancreas bypass surgeries, the risk of which are comparely high and need some advancing technology to assist the procedure. Also, the operations that take a great amount of time, such as cancer, are also costly. Some of the surgeries have high infection rate, such as skin graft, kidney transplant, or burns, the costs of those surgeries will be relatively high.

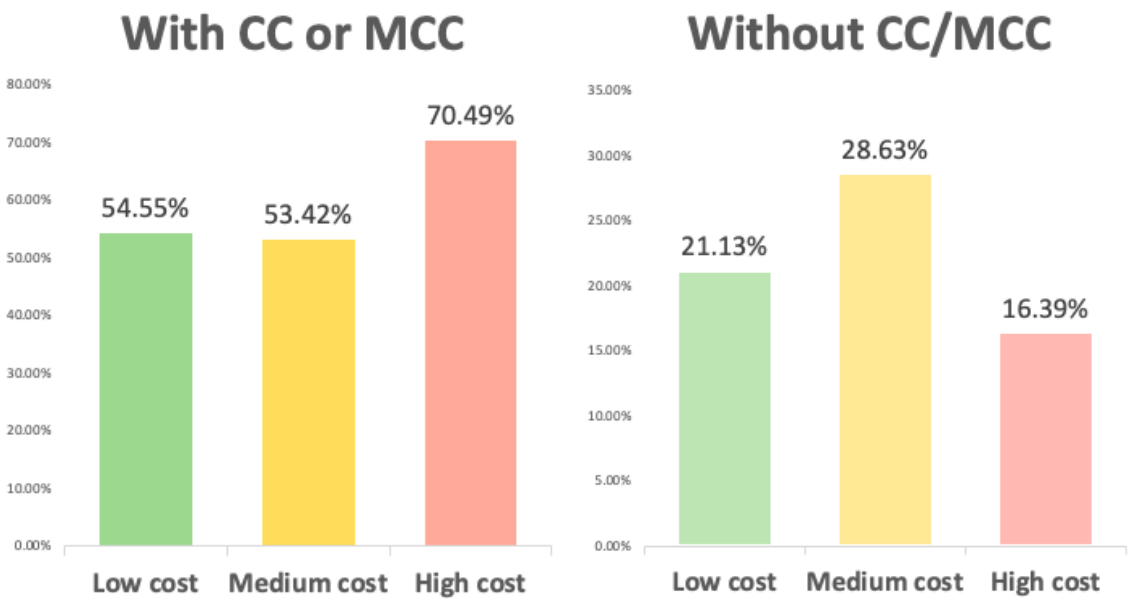
Next, I explore the relationship between MDC Categories and Medium Cost Ratio. The surgeries in medium-cost involve all MDC categories, but only 11 MDC categories are above average. All of these surgeries are not that serve as the operations with high cost ratio. Besides, some of these surgeries have already got mature skills in operating, that's also why its cost under medium status. Apart from these, some chronic diseases are also qualified in this field, because of it long-term low cost of medicine.

Within the low-cost cluster, I observed that all MDC categories have a large low-cost ratio.

The surgeries are not necessary or minor surgeries suffice in this cluster, with some unusual exceptions. Since the operating room and anesthetic are the only CRRC I analyzed, it would be some occasion that I used the room facilities or anesthetic without surgeries. Some examples are mental illness, substance abuse, pregnancy and neonatal. The consultation of physicians makes up a large portion of low-cost DRGs, dealing with the problem of disorder, infection, seizures, disease, and asthma. On the other hand, if an operation is needed, it will probably labelled by another DRG Group with procedure within same MDC category.

I also have some interesting observations that the occurrence of DRG description "with CC", "with MCC", "with CC/MCC" and "without CC/MCC" ("MCC" means "major complication or comorbidity", "CC" means "complication or comorbidity") is extremely high, and some of the DRGs with very similar names has only this part of description in different. Meanwhile, the costs are also sometimes dramatically different among these similar DRGs. Therefore, I calculate the percentage of "with CC MCC" v.s. "Without CC

or MCC" in each cost category. According to the bar charts below,I could find that for DRGs with complication or comorbidity, there are more DRGs fall into high cost than low cost. That makes sense because complication or comorbidity could make operation more complex thus cost more.



In addition, I also analyze the effects of types of services. The services of DRGs are falling into two categories: medical and surgery, while a few of DRGs does not belongs to any of them. Based on the calculations and filters I did on my dataset, I created two charts that explain the findings.

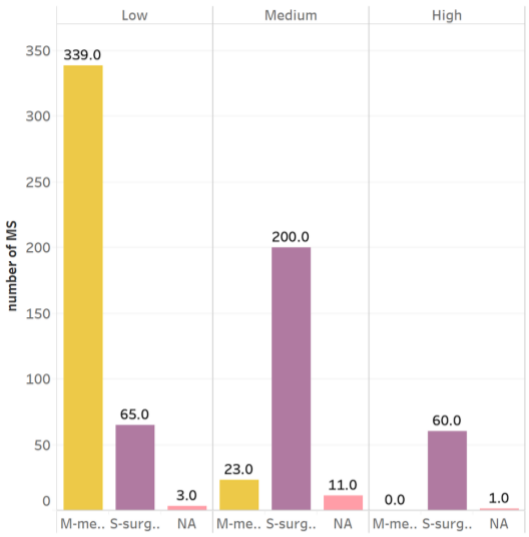


Chart 3.2 number of DRGs per category

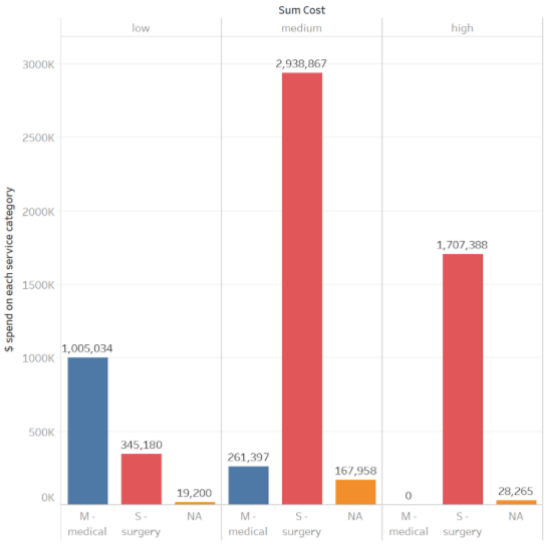


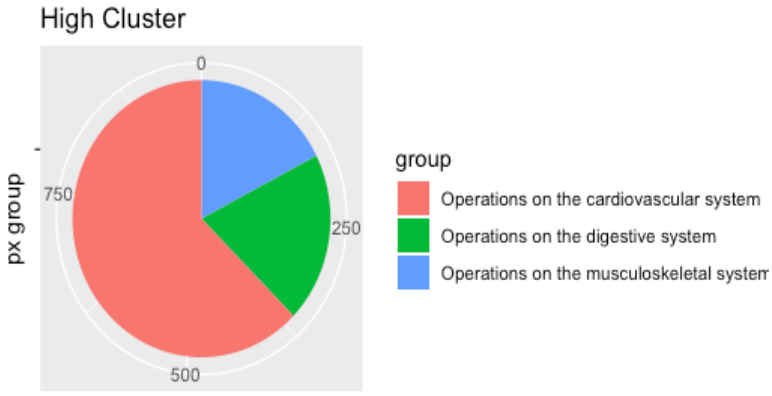
Chart3.1 $ spent per category per cluster

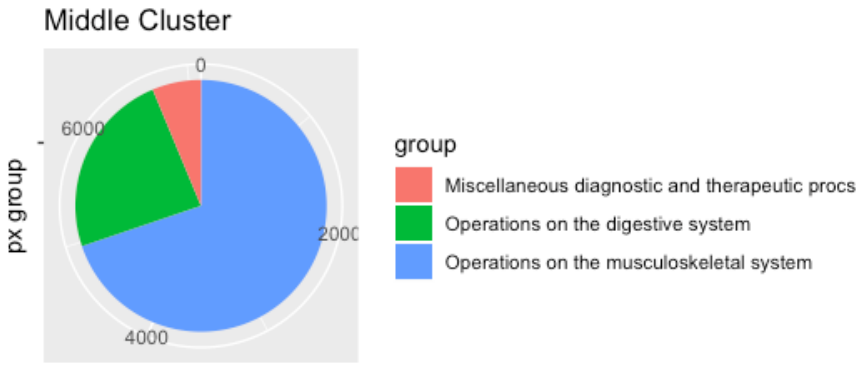
Chart 3.2 represents the number of DRGs per category in each cluster. The number of Medical services makes up most of Low-cost cluster, which is way more than surgery type. 200 out of 234 DRGs are surgery service-type in the medium-cost cluster, and no medical service occur in High cost group. Therefore, when I move from low-cost cluster to high-cost cluster, the number of surgery services increased and the number of medical services decreased accordingly, and the surgery-type domains the high-cost level.

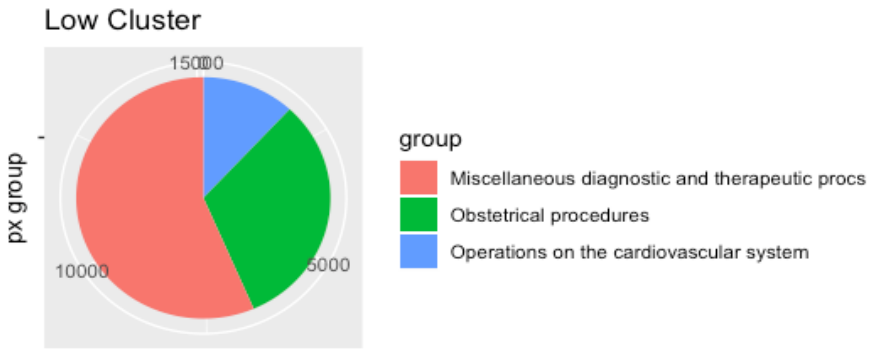
Chart 3.1 shows the dollar spent per category in each cluster. Total costs of Medical services is the highest in the Low-cost cluster, however, the number of medical services is much higher (339) compare with surgeries (65). Therefore, the average cost per DRG in medical category is $2964.7, which is less than the average cost per DRG in surgery ($5310.46). In the medium-cost cluster, surgery costs much more than medical services, and almost all costs fall into the surgery service except one cost with no specific category in the high-cost cluster. Thus, according to the charts, I could conclude that the number of the surgery is an important factor that help with distinguishing clusters. The more surgery contains in the cluster, the higher the cost would be. This makes sense because surgery usually costs much more than a typical medical service, so the clusters with higher k-mean should include more surgery services.

I hope to further confirm my assumption that medical/surgery type may have relationship with the cluster classification.

So, I merge the cluster table with the inpatient table and find the CCSPXGRP categories (CCS PX HIGH LEVEL GROUP) of three clusters. Then I keep the top three PX groups of each cluster for further comparison. From the pie graphs, I notice that High and Medium clusters have more operation PXs, however, high cluster concentrates on cardiovascular system while medium cluster concentrates on the musculoskeletal system. In comparison, low cluster has fewer operations, the major PX group is Miscellaneous diagnostic and therapeutic procs.







Overall, the cluster analysis shows us a picture of hospital DRGs based on operation room usage and anesthetic that the categories of service type are the main factor of cost groups. The higher the cluster mean is, the more surgery services contained in the cluster. Looking into each cluster, I found the High-cost surgeries are about operations on small but important parts, while the medium cost cluster contains operations or medical services on minor parts. The low cost does not contain complex procedures.