Insurance Claims Data & Analytics Part II

Healthcare Data Analytics and Data Mining

Group 3

Jinman Rong
Runxua Cao
Ting Xiao
Xinlan Wu
Yunyi Wang
Ziyu Tang
(Audit) Peihan Tian
(Audit) Fangzhou Xiang

Introduction

This report analyzes the insurance claim data in 2016. Main topics are discussed as follows:

- Study of A Disease Cohort—— Taking Rheumatoid Arthritis as an Example;
- MDC in Inpatient Care Concentration Analysis——A Case Study of MDC 14 Compared with MDC 1;
- Cost Cluster Analysis—— A Case Study of PCCR 3700 Operating Room and PCCR 4000 Anesthesiology.

Through this report, you will get to know how a disease cohort will be conducted, how inpatient care concentration will differentiate among different MDCs, and what factors will contribute to cluster and classify inpatient DRG admissions.

Study of Rheumatoid Arthritis Cohort

Introduction

Cohort studies are a type of medical research used to investigate the causes of disease and to establish links between risk factors and health outcomes. In this part, we will look into Rheumatoid Arthritis (RA) from five perspectives.

Firstly, we mainly identify three most common types of the two kinds of Rheumatoid Arthritis (RA), which are common chronic RA and the other is other Rheumatoid Arthritis with systemic involvement. We also dig into the gender patterns in RA prevalence and test our hypothesis. Furthermore, we analyze the statistics features of patients' total charges and find out the top-5 most common services for treatment of the RA.

Analysis

1. RA Identifications

Every study should initially identify its main object and corresponding symptoms so that researchers can focus on a narrowly defined cohort of patients with certain health conditions to interpret effects of a suspected risk factor.

Below is the table of the most common types of common chronic RA:

Table 1 The three most common types of chronic RA

Code	Dx_Code_Desc	Frequency
M069	Rheumatoid arthritis, unspecified	909
M0579	Rheu arthritis w rheu factor mult site w/o org/sys involv	17
M059	Rheumatoid arthritis with rheumatoid factor, unspecified	8

Unspecified RA is the most frequent type of common chronic RA, which makes sense because this category does not specify the related body parts, with/without rheumatoid factor and with/without involvement with organs and systems. It is the most general type in the cohort. RA with rheumatoid factor in multiple sites without organs and systemic involvement ranks second with the frequency of 17 while RA with rheumatoid factor but no specified sites is the No.3. What is worth to mention is that in this cohort, all categories are not involved with organs and systems.

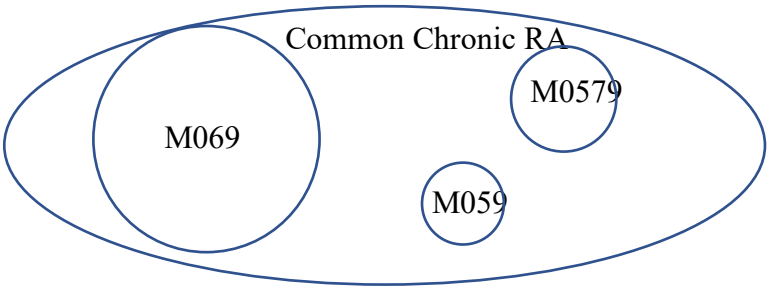


Figure 2 The Venn diagram of the three most common types of chronic RA

As for other Rheumatoid Arthritis with systemic involvement:

Table 2 Other Rheumatoid Arthritis with systemic involvement

Code	Dx_Code_Desc	Frequency
M0510	Rheumatoid lung disease w rheumatoid arthritis of unsp site	21
M0519	Rheumatoid lung disease w rheumatoid arthritis mult site	2
M05671	Rheu arthrit of right ank/ft w involv of organs and systems	2

The frequency of developing Rheumatoid lung disease with rheumatoid arthritis of unspecified site ranks first, followed by Rheumatoid lung disease with rheumatoid arthritis of multiple site and Rheumatoid arthriti of right ankle and foot with involvement of organs and systems with the same frequency of 2.

The result implies that lung disease related with Rheumatoid Arthritis is the most common in the cohort and people with rheumatoid lung disease are more diagnosed with unspecific sites.

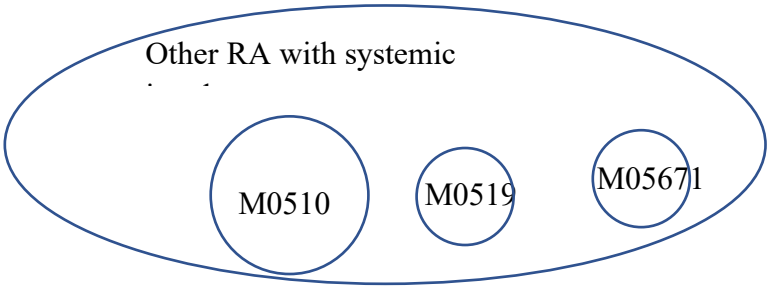


Figure 2 The Venn diagram of other Rheumatoid Arthritis with systemic involvement

2. Gender Analysis

Then, we want to find whether gender will account for RA risk factors. In this part, we will look into whether gender affects the development of common RA or other RA with systemic involvement or RA, separately.

(1) Gender vs RA

When digging into the gender difference in the frequency of developing RA and other RA, the table is as follows:

Table 3 Gender differences in developing RA (all categories)

	RA_freq	Non_RA_freq
Male	267	3367773
Female	745	4043655

Therefore, our null hypothesis is:

H0: Gender does not affect the development of RA or other RA.

After doing Fisher Exact Test, p-value < 2.2e-16. Therefore, we reject the null hypothesis that gender does not affect the development of RA at 5% level. Also, we can conclude from the table that female have higher frequency than male in developing Rheumatoid Arthritis.

(2) Gender vs Common RA

When digging into the gender difference in the frequency of developing RA/other RA and not developing RA/other RA, the tables are as follows:

Table 4 Gender differences in developing common RA

	RA_frequency	Non_RA_frequency
Male	253	3367773
Female	728	4043655

Our first null hypothesis is:

H0: Gender does not affect the development of RA.

After doing Fisher Exact Test, p-value < 2.2e-16 < 0.01. Therefore, we reject the null hypothesis that gender does not affect the development at 1% level. We can also conclude from the table that females have higher frequency than male in developing common Rheumatoid Arthritis.

(3) Gender vs Other RA

Table 5 Gender differences in developing other RA with systemic involvement

	Other_RA_frequency	Non_RA_frequency
Male	14	3367773
Female	17	4043655

Our second null hypothesis is:

H0: Gender does not affect the development of other RA with systemic involvement.

After doing Fisher Exact Test, p-value is 1 > 0.1. Therefore, we fail to reject the null hypothesis that gender does not affect the development of other RA with systemic involvement. As we can conclude from the table that the gender difference between frequencies of other RA with systemic involvement is very small.

In all, Rheumatoid Arthritis is overall a typical disease for female. When we subdivide RA into two sub-categories, we find that females have higher frequency than males in developing common Rheumatoid Arthritis while there is no gender difference in developing other RA with systemic involvement.

3. Variability Analysis

We also calculate four quartiles and IQR of total charges from the revenue code dataset. The IQR is 2757.7, which is four times the first quartile charges, showing a large range of total charges difference.

Table 6 Four Quartiles number

1st Quartile (Min)	2nd Quartile (Median)	3rd Quartile	4th Quartile (Max)
682.5	1521.6	3440.2	227311.8

4. Service Utilization Analysis

Finally, we link two sub-cohort to the Revenue Code and find top five most common services (as defined by the Revenue Codes) for treatment of the RA. Below are the two tables:

Table 7 Top five most common services for RA

REVCODE	Description	RA_frequency
300	Laboratory - Clinical Diagnostic	3222
636	Drugs Require Specific ID: Drugs requiring detail coding	1222
250	Pharmacy	1165
450	Emergency Room	1056
320	Radiology - Diagnostic	375

Table 8 Top five most common services for other RA with Systemic Involvement

REVCODE	Description	Other_RA_frequency
300	Laboratory - Clinical Diagnostic	63
460	Pulmonary Function	48
510	Clinic	18
450	Emergency Room	13
636	Drugs Require Specific ID: Drugs requiring detail coding	13

From frequency tables, we find that Laboratory - Clinical Diagnostic (REVCODE = 300) is the most frequent service in RA in total. Drugs Require Specific ID: Drugs requiring detail coding (REVCODE = 636) is also in top five typical service in both RA and other RA but with a higher rank in RA than other RA. It seems that common services provided for the two main categories have high similarity.

The second basic service for other RA with systemic involvement is pulmonary function, which is in line with the high frequency of rheumatoid lung diseases.

The top service is seemingly to be means of a diagnosis other than treatment. This situation verifies that rheumatoid arthritis has a variety of symptoms and causes, and difficulties to diagnose in its early stages.

In terms of chronic RA, medications are the most common treatment from the tables above.

Comparing two tables, we can conclude that the top five RA services are more frequent than the top five other RA services, which is also in line with the frequency of cohorts analyzed before.

MDC in Inpatient Care Concentration Analysis

——A Case Study of MDC 14 Compared with MDC 1

Introduction

Referral hospitals/centers means the hospitals that can take on any challenge in medicine and operate on virtually any patient no matter how complex the case is. How these referral hospitals/centers distribute is what we try to figure out. To simplify our procedure, we only investigate in MDC 1: Diseases and Disorders of the Brain and Nervous System and MDC 14: Pregnancy, Childbirth and Puerperium these two MDC to represent the complexity of inpatient medical care operation abilities. Calculating the 'lion share' and HHI can help to analyze the concentration of the market and draw a picture of those big players alike.

Guess Before Analysis

We think that MDC 1: Diseases and Disorders of the Brain and Nervous System will be more concentrated because it involves exploring nervous system and brain functioning. And this is an area that humans are still exploring, and thus requires more talented people and sophisticated equipment. Meanwhile, MDC 14: Pregnancy, Childbirth and Puerperium is much more like to be included in common symptoms. It may not require that precise equipment.

HHI Index Calculation

MDC 14: Pregnancy, Childbirth and Puerperium

The operation in MDC 1: Diseases and Disorders of the Brain and Nervous System means the complexity that the medical centers can conduct, while the MDC 14: Pregnancy, Childbirth and Puerperium shows the comprehensive of them. The number of patients counts, and total charge money will be compared in these vertical and horizontal level by calculating each medical centers' share, so we can get the following tables:

Table 9 Market concentration of MDC 14 calculated by patient counts

Hnum2	HOSP_DESC	Total	Share
5	University of Vermont Medical Center (as of 2014)	2416	0.425877
16	Southwestern Vermont Medical Center	471	0.083025
9	Porter Medical Center	379	0.066808
1	Northwestern Medical Center	378	0.066631
8	Rutland Regional Medical Center	361	0.063635
6	Central Vermont Hospital	326	0.057465
15	Brattleboro Memorial Hospital	306	0.053940
3	Northeastern Vermont Regional Hospital	226	0.039838
2	North Country Hospital and Health Center	206	0.036312
4	Copley Hospital	206	0.036312
10	Gifford Memorial Hospital	205	0.036136
12	Springfield Hospital	193	0.034021

Table 10 Market concentration of MDC 14 calculated by total charges

Hnum2	HOSP_DESC	Total	Share
5	University of Vermont Medical Center (as of 2014)	30773062.5	0.465255
16	Southwestern Vermont Medical Center	4907476.72	0.074196
9	Porter Medical Center	4824498.38	0.072941
8	Rutland Regional Medical Center	4171436.69	0.063068
6	Central Vermont Hospital	4049609.19	0.061226
15	Brattleboro Memorial Hospital	3053784.51	0.04617
10	Gifford Memorial Hospital	2947250.75	0.044559
3	Northeastern Vermont Regional Hospital	2845266.63	0.043017
2	North Country Hospital and Health Center	2528758.28	0.038232
1	Northwestern Medical Center	2384525.14	0.036051
12	Springfield Hospital	2144511.09	0.032423
4	Copley Hospital	1512193.61	0.022863

MDC 1: Diseases and Disorders of the Brain and Nervous System

Same tables are made for MDC 1:

Table 11 Market concentration of MDC 1 calculated by patient counts

Hnum2	HOSP_DESC	Total	Share
5	University of Vermont Medical Center (as of 2014)	2039	0.622405
8	Rutland Regional Medical Center	335	0.102259
6	Central Vermont Hospital	204	0.062271
16	Southwestern Vermont Medical Center	135	0.041209
1	Northwestern Medical Center	111	0.033883
10	Gifford Memorial Hospital	85	0.025946
2	North Country Hospital and Health Center	62	0.018926
12	Springfield Hospital	62	0.018926
3	Northeastern Vermont Regional Hospital	60	0.018315
15	Brattleboro Memorial Hospital	57	0.017399
4	Copley Hospital	55	0.016789
9	Porter Medical Center	41	0.012515
11	Mount Ascutney Hospital and Health Center	15	0.004579
14	Grace Cottage Hospital	15	0.004579

Table 12 Market concentration of MDC 1 calculated by total charges

Hnum2	HOSP_DESC	Total	Share
5	University of Vermont Medical Center (as of 2014)	92200270	0.792715
8	Rutland Regional Medical Center	9202322	0.079119
6	Central Vermont Hospital	3813673	0.032789
16	Southwestern Vermont Medical Center	2219602	0.019084
10	Gifford Memorial Hospital	1604507	0.013795
3	Northeastern Vermont Regional Hospital	1369112	0.011771
1	Northwestern Medical Center	1287275	0.011068

2	North Country Hospital and Health Center	1069212	0.009193
12	Springfield Hospital	1015393	0.00873
15	Brattleboro Memorial Hospital	955791.7	0.008218
9	Porter Medical Center	715206.9	0.006149
4	Copley Hospital	448450.6	0.003856
11	Mount Ascutney Hospital and Health Center	220861	0.001899
14	Grace Cottage Hospital	187742.6	0.001614

After calculating, we can find that both in MDC 1 and MDC 2, the University of Vermont Medical Center has the 'lion share' in patient numbers and charges.

However, the significance of 'lion share' only has meanings in the highly concentrated market, so we want to first take a look at the market structure by calculating 0-1 HHI.

Table 13 Market structure by calculating 0-1 HHI

MDC-14		MDC-1	
HHI-Counts	HHI-$	HHI-Counts	HHI-$
0.214	0.465	0.407	0.637

As HHI in MDC-1 is larger than 0.3, while MDC-14 is smaller than 0.3, we can conclude that the MDC-1 is a highly concentrated market. The 'lion share' in MDC-1 is the University of Vermont Medical Center, and its share is 0.79 in patients' charges and 0.622 in patients' numbers.

Lion Share Hospital Description

From the analysis above, the University of Vermont Medical Center seems to be the 'lion' for both MDC-1 and MDC-14. Its shares are shown below:

Table 14 Market Share of lion 'the University of Vermont Medical Center'

MDC-14 Share		MDC-1 Share	
Counts	Total charges-$	Counts	Total charges-$
0.426	0.465	0.622	0.792

Let's have a close look at the University of Vermont Medical Center. Vermont Medical Center provides advanced care to approximately one million people in Vermont and northern New York as regional referral center and to approximately 160,000 residents in the Chittenden and Grand Isle Vermont counties as a community hospital. That's may be the reason why it has a huge patient number among all other medical center. Also, the hospital offers the region advanced technology and techniques. 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.

Summary

In conclusion, most complicated surgical procedures involving aspects of the internal nervous system and brain functioning may concentrate on some high technology medical centers, like MCD-14, in this part. Thus, our guess before the calculation is correct. However, we only compare the MCD-1 and MCD-14 to test the ability to conduct a complex operation without considering other surgery. Also, the influence of the population in the neighborhood to the medical center's patient number is not excluded. Here may be individual cases so that further research needed to drive a final conclusion.

Cost Cluster Analysis

Data Cleaning & Manipulating Process Description

We first filtered hospital admissions to only important DRGs between 20 and 977 and dropped the revenue charges which are less than 100. Then, we converted the code of DRG and PCCR to the names and made a cross table with the DRGs in the row and mean value of the PCCRs, consisting of 687 rows and 54 columns. We created a new variable called PCCR_OR_and_Anesth_Costs which is the combination of PCCR 3700 and PCCR 4000, turned all empty cells to zero and used these values for as costs for clustering.

Cluster Analysis

For the cluster analysis, we first try to cluster the cost data into 2,3,4 and 5 clusters and examine the Calinski-Harabasz f-statistics to see the best clustering of cost. The result is shown as below.

Table 35 Calinski-Harabasz f-statistics of different numbers of clusters

Number of clusters	f-statistics
2	1432.768
3	1700.194
4	1934.762
5	2134.944

It can be seen that the more clusters the higher f-statistics, indicating that 5 clusters may be the best clustering of costs. However, for a more convenient and intuitive explanation, we only focus on solutions with 3 clusters.

By using K-Means algorithm, we got three groups of DRGs, the low-cost group with the mean cost of $2,841 and size of 397, the medium-cost group with the mean cost of $14,143 and size of 229, and the high-cost group with the mean cost of $28,339 and size of 61. The graph of three clusters of DRGs is shown as below.

It can been seen that the low-cost DRG is more common to see than the high-cost ones. Furthermore, the low-cost group has many observations of 0, indicating that many DRGs of the low-cost group doesn't need operating room or anesthesiology.

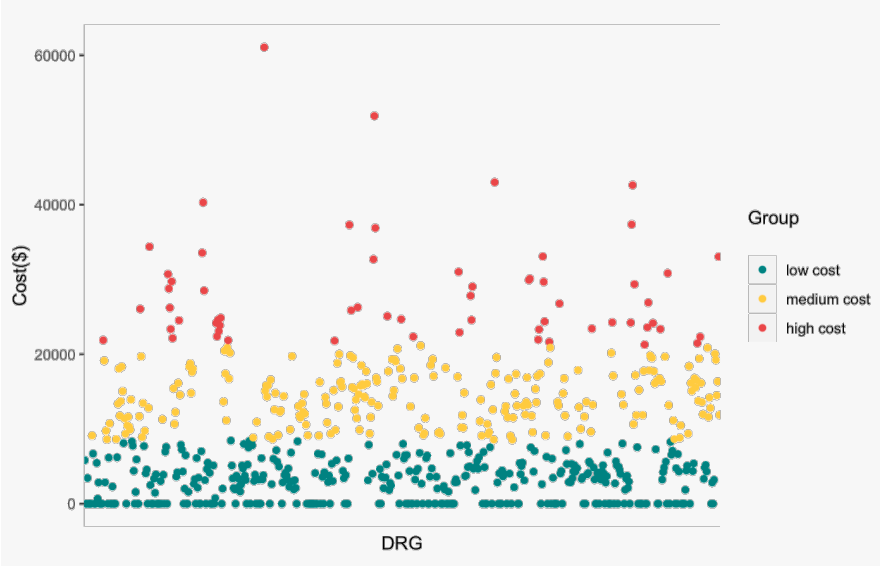


Figure 3 Three clusters of DRGs

Why the DRGs of three clusters are relatively similar related to one another with in the group but relatively dissimilar to other clusters' DRGs? The reasons are the different properties of the DRGs, including the seriousness of DRGs, the main treatments needed by DRGs, and the concentration of the care for DRGs.

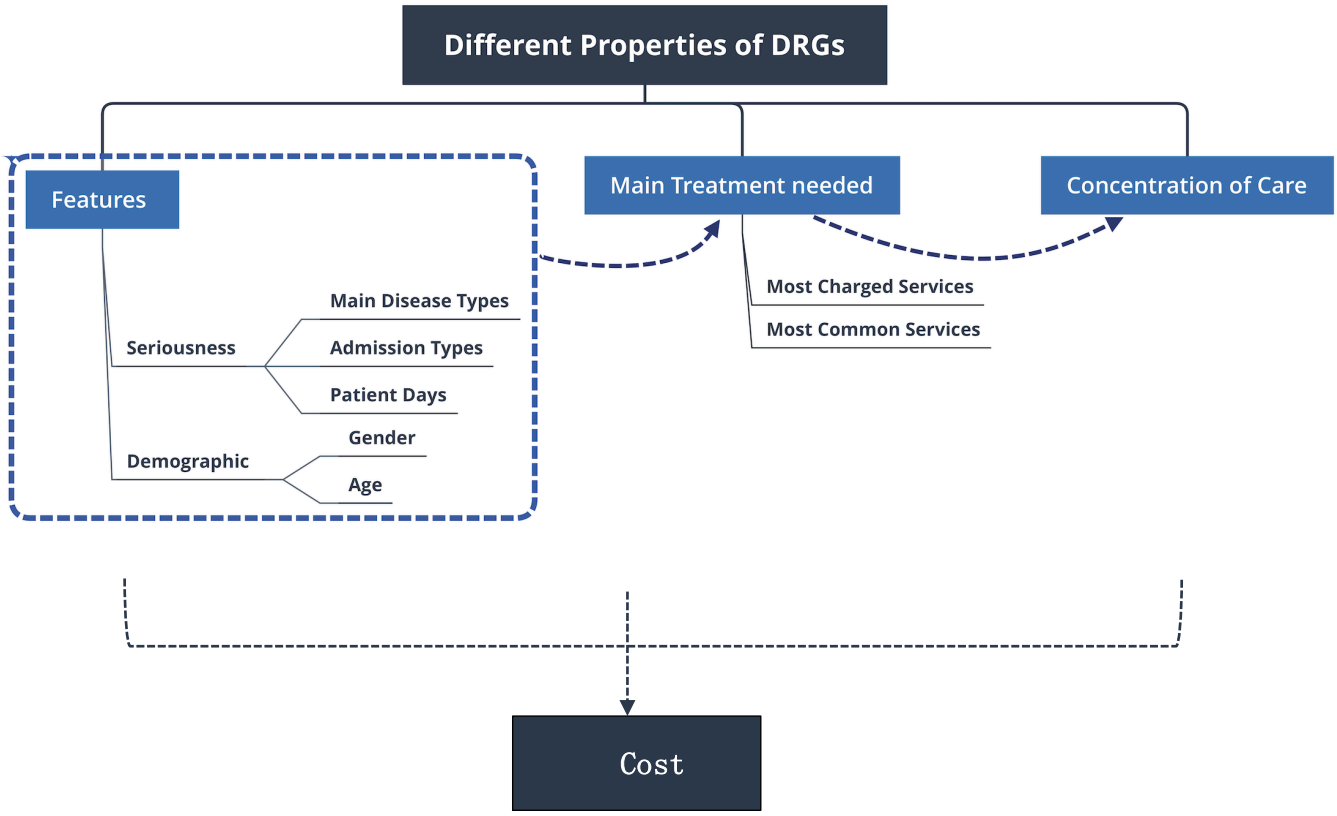


Figure 4 Analysis logic map

The seriousness of DRGs determines the main treatments they need, finally resulting in the different concentration of the care for them, which are reflected in terms of the cost.

The Seriousness of DRGs

We use three features to show the seriousness of DRGs among three groups: main disease types, admission type, and patient days.

1. Main Disease Types

We apply word cluster techniques by counting DRG description frequencies to see what symptoms or services are common in a particular group.

(1) Low-cost Group

The word cluster for low-cost group is show as follows:

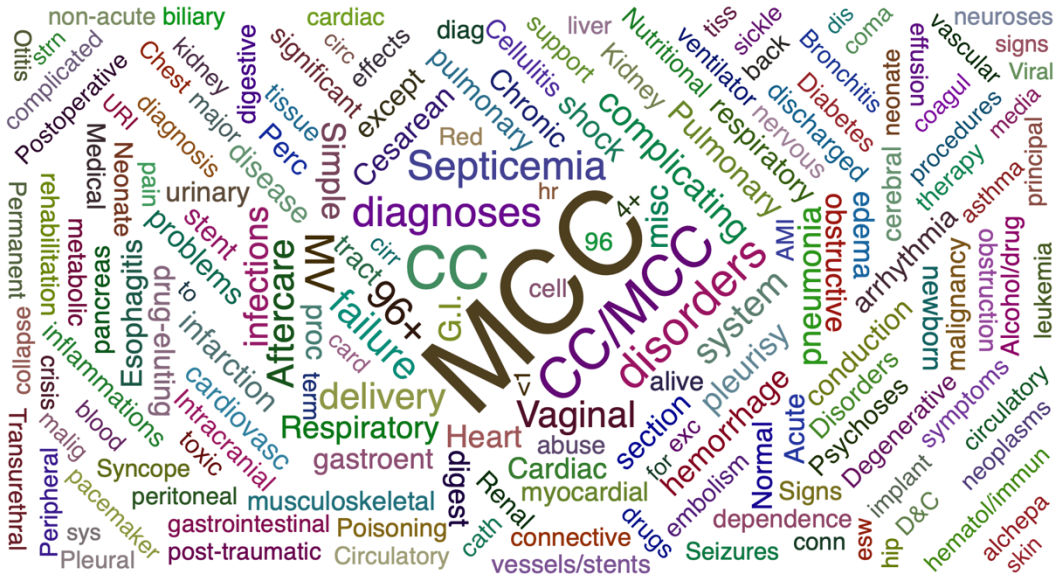


Figure 5Low-cost Group Word Cloud

As we can see from the graph, this group involves mostly common surgical procedures, which are of less risk and modern medicine could be able to basically solve most problems.

(2) Medium-cost Group

The word cluster for medium-cost group is show as follows:

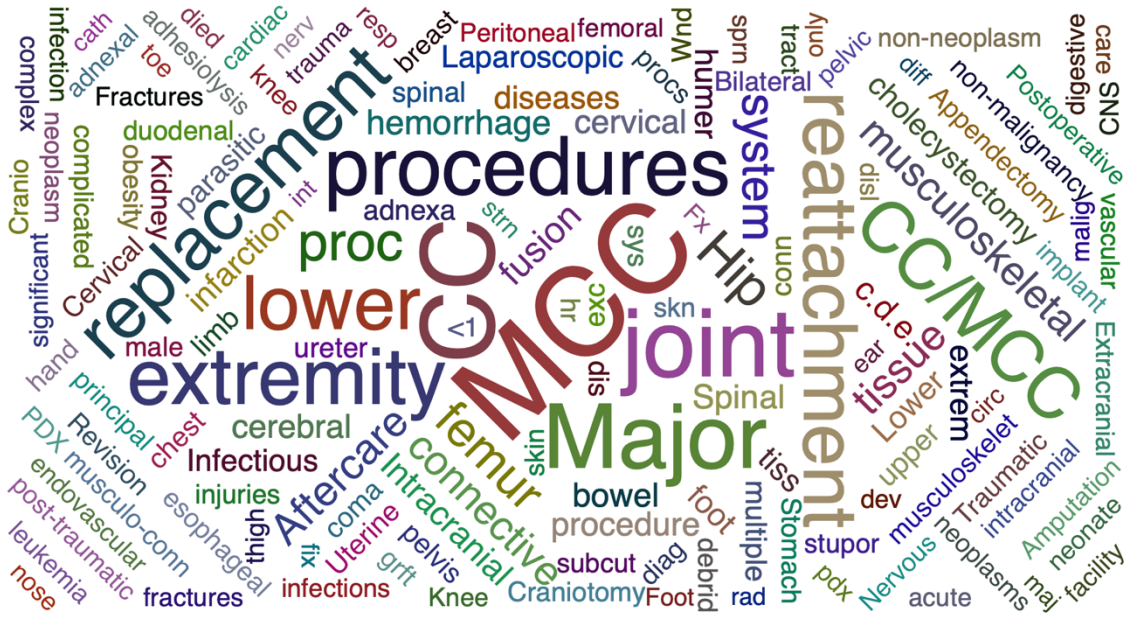


Figure 6 Medium-cost Group Word Cloud

This group mostly consists of external damages, such as extremity and femur. These kinds of disease may not be deadly but could occur more frequently and need more time to recover.

(3) High-cost Group

The word cluster for high-cost group is show as follows:

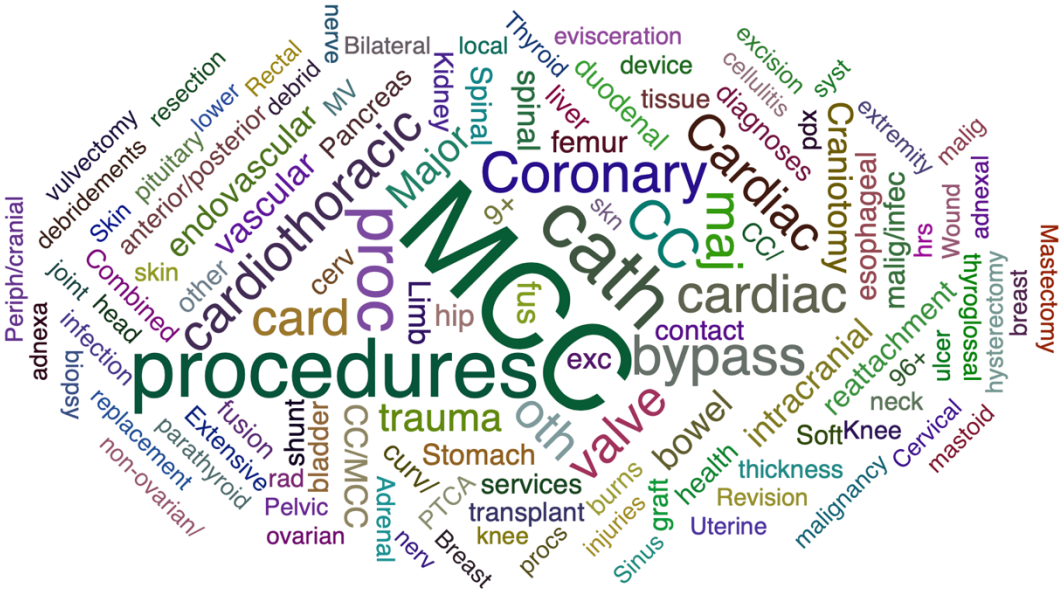
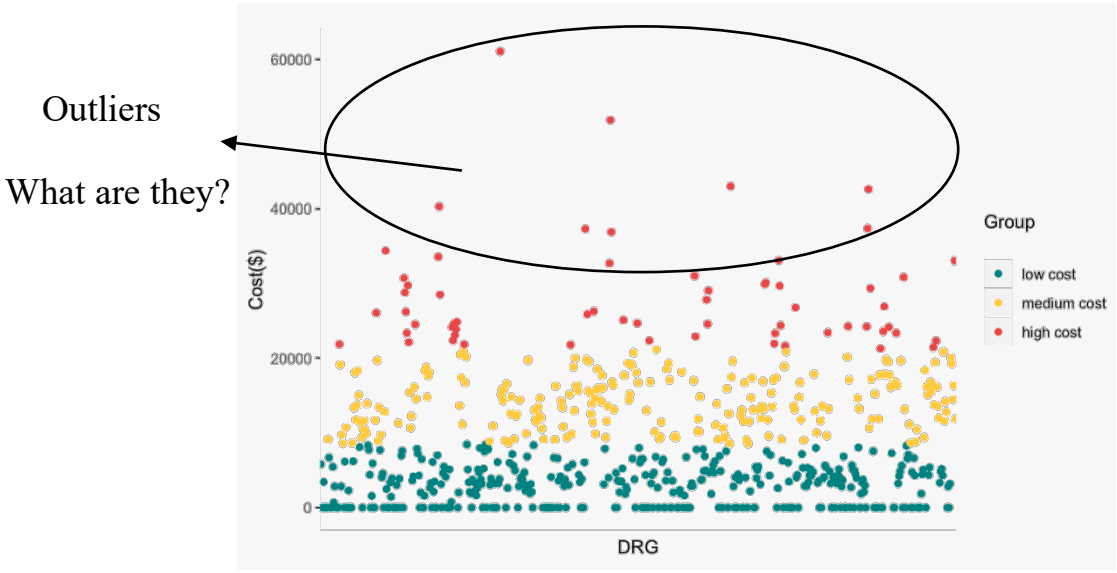


Figure 7 High-cost Group Word Cloud

Unlike the words above, in this group, we can see high frequency in words like cardiothoracic, coronary, etc., which are quite severe diseases. They may require complex surgical and external maintainers. Thus, they tend to be much expensive than the other two.

Outliers Analysis



Then, we filter these outliers by sorting them using cost. The table is presented as follows:

Table 16 Descriptions and Costs of Outliers

DRG_Descriptions	PCCR_OR_and_Anesth_Costs ($)
Extensive burns or full thickness burns w MV 96+ hrs w skin graft	61064.86
Major bladder procedures w MCC	51883.38
Other heart assist system implant	43012.84
Skin graft &/or debrid exc for skin ulcer or cellulitis w MCC	42624.37
Combined anterior/posterior spinal fusion w MCC	40307.15
Skin graft &/or debrid exc for skin ulcer or cellulitis w CC	37365.43
Kidney transplant	37310.69
Major bladder procedures w/o CC/ MCC	36897.68
Breast biopsy, local excision & other breast procedures w/o CC/MCC	34389.78

The table further verifies that diseases related to skin or cardiovascular disease, which are correlated to live transplants, along with long-time observations, are categorized together. The table also reveals gender characteristics, such as breast surgery mostly for women. We will analyze more in the gender analysis.

In all, after preliminary analysis, we can see that surgeries related to cardiovascular disease or other disease-related to internal organs, which required precision instruments, are categorized as one group. In addition, common operations involved in trauma with a lower probability of death are grouped. Finally, contusion of various tissues of the body and other related diseases are considered as one group.

2. Admission Type

From the perspective of admission types, the most common admission type of low-cost group is emergency. Combined with the results of words cloud that the most common diseases are the less severe injuries, we can imagine the scenario of these DRGs of low-cost group: a person comes to emergency room with an injury that doesn't kill the life. By contraries, the most common admission type of medium-cost and high-cost groups is elective. Combined with the results of words cloud that the diseases of these two groups are more severe than the low-cost ones, people need to go to the special departments of the hospital for certain diagnosis and treatment rather than the emergency room.

3. Patient Days

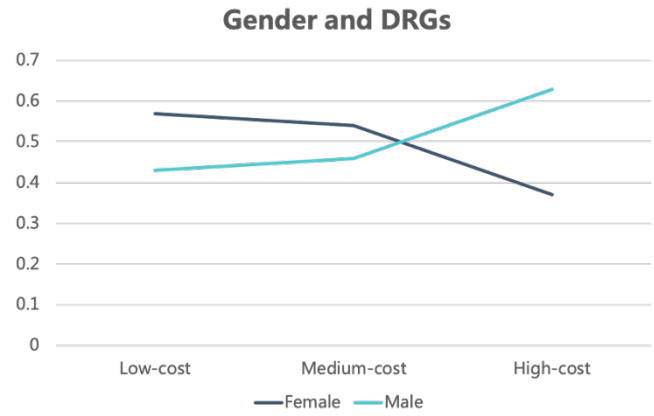
The patient days also reflect the seriousness of the DRGs among different groups. The average patient days among low-cost, medium-cost, and high-cost groups is shown as below. It's obvious that the patient days of high-cost group is significantly more than that of low-cost and medium-cost, indicating that high-cost group has the most severe DRGs and this is why the DRGs of high-cost group costs most among three groups.

Table 17 Average patient days among three groups

Groups	Average patient days
low-cost group	4.3
medium-cost group	4.7
high-cost group	9.5

Demographic Statistics

Demographic Statistics is applied here to explore whether it will help cluster different groups of DRGs.



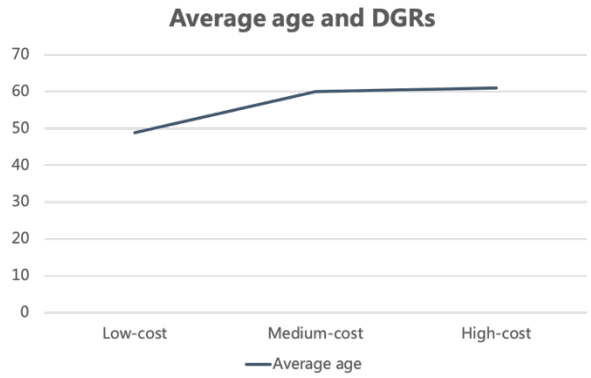


Figure 8 Lines of gender and DRGs relationships Figure 9 Line of average age and DRGs relationships

In terms of gender, females are more likely to involved in low-cost and medium-cost groups, while males much tend to develop symptoms in the high-cost category. The gap is widened when it comes to the high-cost type.

The average age graph shows a slightly upward trend and later becomes almost horizontal. In reality, it is indeed for older people to break legs because of osteoporosis or have cardiovascular diseases.

Thus, gender and age both could contribute to the interpretation of cluster.

The Main Treatments Needed by DRGs

The seriousness of DRGs determines what kind of medical services and treatments are needed and how much it will cost. We analyzed the top 5 medical services in terms of total charges and total units among three groups. The main treatments needed serves a meaningful feature that varies among three groups. The results are shown as below.

Table 18 The top 5 most charged medical services among three groups

low-cost group	medium-cost group	high-cost group
Room & Board (Semi-Private 2 beds)	Operating Room Services	Operating Room Services
Laboratory - Clinical Diagnostic	Medical/Surgical Supplies: Other implants	Room & Board (Semi-Private 2 beds)
Emergency Room	Room & Board (Semi-Private 2 beds)	Medical/Surgical Supplies: Other implants
Pharmacy	Pharmacy	Intensive care
Psychiatric	Anesthesia	Pharmacy

Table 19 The top 5 most common medical services among three groups

low-cost group	medium-cost group	high-cost group
Pharmacy	Pharmacy	Pharmacy
Drugs Require Specific ID: Drugs requiring detail coding	Anesthesia	Drugs Require Specific ID: Drugs requiring detail coding
Pharmacy: Other	Operating Room Services	Operating Room Services
Drugs Require Specific ID: Self admin drugs	Drugs Require Specific ID: Drugs requiring detail coding	Anesthesia
Laboratory - Clinical Diagnostic	Drugs Require Specific ID: Self admin drugs	Drugs Require Specific ID: Self admin drugs

We find that the services needed are general with relatively low cost in the low-cost group. The most charged service of low-cost group is Room & Board, however the patients with the diseases in this group won't stay in hospital

for a long time, indicating that there won't be too much charges compared with the other two groups. Furthermore, the top 5 common services are most about the drugs, supposing that the treatments for DRGs in the low-cost group is less technical and can be done by most hospitals.

As for medium-cost and high-cost group, operating room service and anesthesiology are deadly needed not only in term of total charges, but also total units. This is also consistent with the fact that the costs of combination of operating room service and anesthesiology of these two groups are relatively higher. The services needed by medium-cost and high-cost group is more complex than the ones in low-cost group, such as surgery. Especially, the high-cost group needs intensive care which is expensive and not provided by all medical centers, indicating that high-cost group has the most severe DRGs, needs the most advanced and complex treatments which may not be done by all hospitals.

The Concentration of the Care for DRGs

The concentration of the care for DRGs is another important feature as it is different among three groups due to the distinct needs of treatments. We calculated the market shares of total charges and patient counts of the hospitals among three groups and used HHI index to show the concentration of the care. The result is shown as below.

Table 20 The concentration of the care for DRGs among three groups

	HHI of total charges	HHI of patient counts
low-cost group	0.189	0.182
medium-cost group	0.259	0.254
high-cost group	0.760	0.755

It can be concluded that nearly all the hospitals are willing and be able to perform medical care in the DRGs of low-cost group, in other words, the DRGs of low-cost group can be done more generally by most of the hospitals, making the HHI of total charges and HHI of patient counts are the lowest among the three.

By contrary, due to the seriousness and complexity of DRGs of high-cost group, they tend to be highly concentrated among specialized high technology medical centers such as the University of Vermont Medical Center, making the HHI of total charges and HHI of patient counts highest among the groups. Since only few hospitals or medical centers are well equipped with technology and medical staff and can take on any challenge, the patients with the diseases of high-cost group will tend to go to these advanced hospitals for high-quality medical services.

The HHI index also explains the reason why the cost of high-cost group is significantly higher than the low-cost group. When a medical service is in a state of monopoly, whether because of the complexity of technology or limitation of availability, according to the principle of market economy, its cost will definitely be higher.

Summary

The first section takes Rheumatoid Arthritis as an example to conduct a disease cohort study. It turns out that females are more likely to have RA and variability analysis, and service utilization analysis shows that rheumatoid arthritis has a variety of symptoms and causes, and difficulties to diagnose in its early stages.

The second section assumes and analyzes inpatient care concentration between MCD 1 and MCD 14. Most complicated surgical procedures involving aspects of the internal nervous system and brain functioning may concentrate on some high technology medical centers.

The third section aims to do cluster analysis and interpret why certain types of DRGs are clustered together. Although five numbers of clusters turn out to be the best, we try to simplify the problem with 3 clusters. We interpret the issue from three perspectives, namely, features of DRGs, the main treatments needed by DRGs, and the concentration of care for DRGs, and these factors all more or less contribute to the formation of clusters.

Appendix

Section 1

```
# load datasets
outpatient      <- read.csv("~/Downloads/VTOUTP16.TXT", header = TRUE)
revenue_code    <- read.csv("~/Desktop/Health analysis and data mining/VTREVCODE16.TXT", header = TRUE)
RA_ICD10_Codes1 <- read_xlsx("~/Downloads/RA_ICD10_Codes.xlsx", sheet = "rheumatoid arthritis")
RA_ICD10_Codes2 <- read_xlsx("~/Downloads/RA_ICD10_Codes.xlsx", sheet = "Other RA w Systemic Invlvmnt")
# tidy datasets: RA_ICD10_Codes1 and RA_ICD10_Codes2
RA_ICD10_Codes1_new = subset(RA_ICD10_Codes1, select = c("ICD-10 Codes", "Dx_Code_Desc"))
RA_ICD10_Codes2_new = subset(RA_ICD10_Codes2, select = c("ICD-10 Codes", "Dx_Code_Desc"))
# tidy datasets: outpatient
data_long           <- gather(outpatient, type, code, DX1:DX20, factor_key = TRUE)
outpatient_selected =  subset(data_long, select = c("Uniq", "sex", "code"))
outpatient_selected = as.data.table(outpatient_selected)
RA                  =  merge(outpatient_selected, RA_ICD10_Codes1, by.x = "code", by.y = "ICD-10 Codes")
other_RA            =  merge(outpatient_selected, RA_ICD10_Codes2, by.x = "code", by.y = "ICD-10 Codes")
RA                  =  subset(RA, select = code:Dx_Code_Desc)
other_RA            =  subset(other_RA, select = code:Dx_Code_Desc)
No_RA               = anti_join(outpatient_selected, RA, by = "code")
No_RA_Other_RA      = anti_join(No_RA, other_RA, by = "code")
# get the frequency of each RA ICD-10 code in each sub-cohort
RA_grouped       <- group_by(RA, code, Dx_Code_Desc)
RA_freq          =  as.data.table(summarise(RA_grouped, frequency = n()))
RA_ordered       =  setorder(RA_freq, -"frequency")
head(RA_ordered,3)
write.csv(head(RA_ordered,3), "~/Downloads/RA_ordered.csv")
other_RA_grouped <- group_by(other_RA, code, Dx_Code_Desc)
other_RA_freq    =  as.data.table(summarise(other_RA_grouped, frequency = n()))
other_RA_ordered =  setorder(other_RA_freq, -"frequency")
head(other_RA_ordered,3)
write.csv(head(other_RA_ordered,3), "~/Downloads/other_RA_ordered.csv")
# identify gender differences in each sub-cohort(Fisher's Exact Test)
RA_grouped_sex         = group_by(RA, sex)
RA_freq_sex            = as.data.table(summarise(RA_grouped_sex, RA_freq = n()))
other_RA_grouped_sex   = group_by(other_RA, sex)
other_RA_freq_sex      = as.data.table(summarise(other_RA_grouped_sex, other_RA_freq = n()))
No_RA_Other_RA_grouped_sex = group_by(No_RA_Other_RA , sex)
No_RA_Other_RA_freq_sex = as.data.table(summarise(No_RA_Other_RA_grouped_sex, No_RA_Other_RA_freq = n()))
No_RA_Other_RA_freq = drop_na(No_RA_Other_RA_freq_sex)
No_RA_Other_RA_freq$No_RA_Other_RA_freq

total1 = cbind(RA_freq = RA_freq_sex$RA_freq, No_RA_Other_RA_freq =No_RA_Other_RA_freq$No_RA_Other_RA_freq)
total_matrix1          = as.matrix(total1)
```

```
rownames(total_matrix1) = c(1,2)
total_matrix1
fisher.test(total_matrix1)
total2 = cbind(other_RA_freq = other_RA_freq_sex$other_RA_freq, No_RA_Other_RA_freq
=No_RA_Other_RA_freq$No_RA_Other_RA_freq)
total_matrix2          = as.matrix(total2)
rownames(total_matrix2) = c(1,2)
total_matrix2
fisher.test(total_matrix2)
# calculate 4 quartiles and IQR of CHRGS
summary(outpatient$CHRGS)
IQR(outpatient$CHRGS)
# Link two sub-cohort to the Revenue Code
RA_RC                =  merge(RA, revenue_code, by = "Uniq")
RA_RC_selected       =  subset(RA_RC, select = c("Uniq", "code", "Dx_Code_Desc", "REVCODE"))
RA_RC_grouped        <- group_by(RA_RC_selected, REVCODE)
RA_RC_freq           =  as.data.table(summarise(RA_RC_grouped, RA_frequency = n()))
RA_RC_ordered        =  setorder(RA_RC_freq, -"RA_frequency")
head(RA_RC_ordered,5)
write.csv(head(RA_RC_ordered,5), "~/Downloads/RA_RC_ordered.csv")
other_RA_RC          =  merge(other_RA, revenue_code, by = "Uniq")
other_RA_RC_selected =  subset(other_RA_RC, select = c("Uniq", "code", "Dx_Code_Desc", "REVCODE"))
other_RA_RC_grouped  <- group_by(other_RA_RC_selected, REVCODE)
other_RA_RC_freq     =  as.data.table(summarise(other_RA_RC_grouped, other_RA_frequency = n()))
other_RA_RC_ordered  =  setorder(other_RA_RC_freq, -"other_RA_frequency")
head(other_RA_RC_ordered,5)
write.csv(head(other_RA_RC_ordered,5), "~/Downloads/other_RA_RC_ordered.csv")
```

Section 2

```
inpatient=inpatient[,c('hnum2','CHRGS','UNIQ','MDC')]
MDC1=inpatient[MDC==1]
MDC14=inpatient[MDC==14]
total1_p=sqldf('select hnum2,count(hnum2) as total from MDC1 group by hnum2')
total1_p=as.data.table(total1_p)
total1_p[,share:=total/nrow(MDC1)]
HHI_1_patient=sqldf('select sum(share*share) from total1_p')
HHI_1_patient
write.csv(total1_p,file='C:/Users/xiaot/Desktop/3.csv')
#
total14_p=sqldf('select hnum2,count(hnum2) as total from MDC14 group by hnum2')
total14_p=as.data.table(total14_p)
total14_p[,share:=total/nrow(MDC14)]
HHI_14_patient=sqldf('select sum(share*share) from total14_p')
```

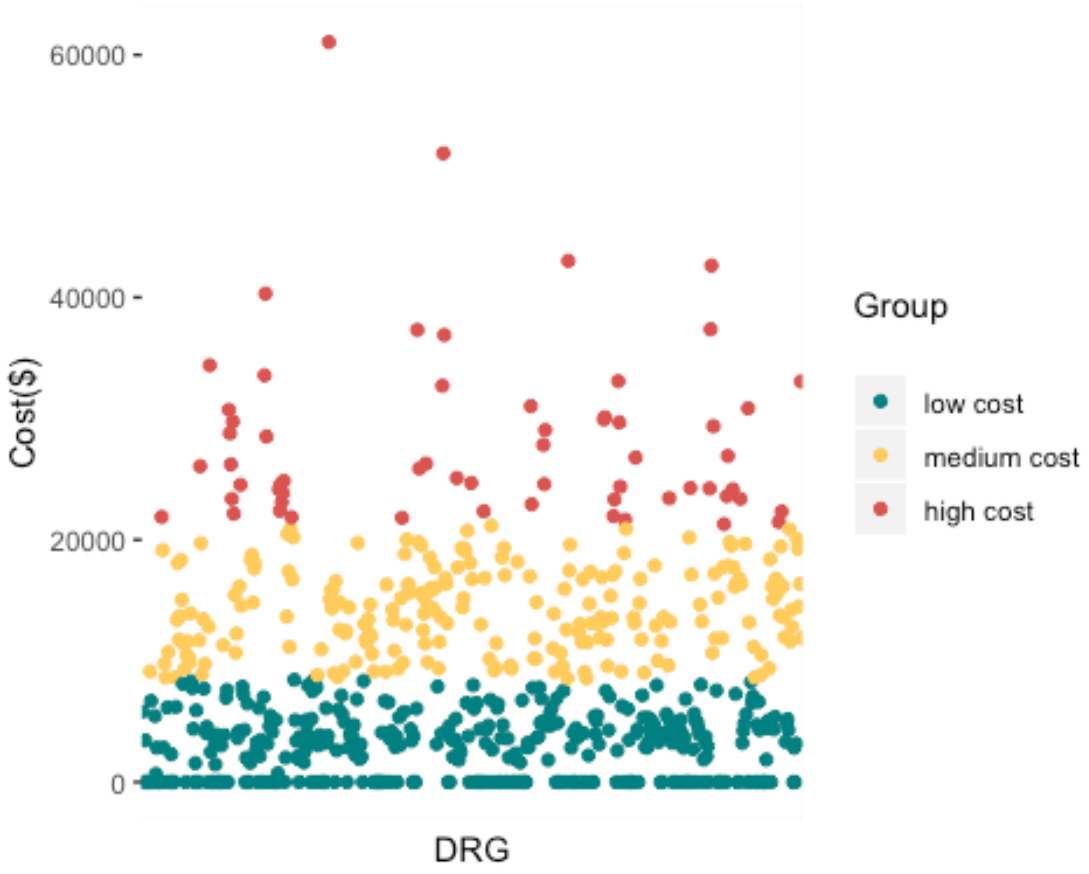
```
HHI_14_patient
total1_c=sqldf('select hnum2,sum(CHRGS) as total from MDC1 group by hnum2')
total1_c=as.data.table(total1_c)
total1_c[,share:=total/MDC1[,sum(CHRGS,na.rm=T)]]
HHI_1_charge=sqldf('select sum(share*share) from total1_c')
HHI_1_charge
total14_c=sqldf('select hnum2,sum(CHRGS) as total from MDC14 group by hnum2')
total14_c=as.data.table(total14_c)
total14_c[,share:=total/MDC14[,sum(CHRGS,na.rm=T)]]
HHI_14_charge=sqldf('select sum(share*share) from total14_c')
HHI_14_charge
```

Section 3

```
 remove(list = ls())
library(data.table)
library(sandwich)
library(tidyverse)
library(lmtest)
library(ggplot2)
library(knitr)
library(psych)
library(dplyr)
library(tidyr)
library(scales)
library(RColorBrewer)
library(reshape)
library(clusterSim)
library(wordcloud)

inp=fread(file="~/Downloads/VTINP16_upd.TXT")
revcode=fread(file="~/Downloads/VTREVCODE16.TXT")
rev=revcode[,c(5,7,10)]
DRG=fread(file="~/Downloads/DRG.csv")
PCCR=fread(file="~/Downloads/PCCR.csv")
inp_new=fread(file="~/Downloads/VTINP16_upd.TXT",
          col.names=c('Uniq','DRG'),
          select = c('UNIQ','DRG'))
inp_new=inp_new[DRG>=20&DRG<=977]

inp_rev=merge(inp_new,rev,by="Uniq",all.x = FALSE)
inp_rev=inp_rev[REVCHRGS>100|REVCHRGS==100]
charge_ad=inp_rev[,.(charge=sum(REVCHRGS,na.rm=TRUE)),keyby =.(Uniq,DRG,PCCR)]
charge_ad=charge_ad[!is.na(PCCR)]
```

```
charge_ad=merge(charge_ad,DRG,by="DRG",all.x=TRUE)
charge_ad=merge(charge_ad,PCCR,by="PCCR",all.x=TRUE)
charge_ad=charge_ad[,-c(1,2)]
colnames(charge_ad) <- c("Uniq", "charge","DRG","PCCR")

total_charge=charge_ad[,.(num_ad=.N,total=sum(charge,na.rm=TRUE)),keyby =.(DRG,PCCR)]
avg_charge=total_charge[,.(average=total/num_ad),keyby =.(DRG,PCCR)]

table=cast(avg_charge,DRG~PCCR)

## Using average as value column.  Use the value argument to cast to override this choice

table=table[,-56]
table=table[-688,]
rownames(table)=table$DRG
table_new=table[,-1]

table_new$PCCR_OR_and_Anesth_Costs=table_new[,"Operating Room"]+table_new[,"Anesthesiology"]
table_new[is.na(table_new)]=0
#table_new
#write.csv(table_new, file="~/Downloads/table_new.csv")

cluster_data=table_new[,55]
cost=as.data.frame(cluster_data)
rownames(cost)=table$DRG
check=cost
check$name=rownames(cost)
check=as.data.table(check)

# Calinski-Harabaz f-stat
set.seed(123456)
for (i in 2:5) {
  result=kmeans(cost,i)
  print(index.G1(cost,result$cluster,centrotypes="centroids"))
}

## [1] 1432.768
## [1] 1700.194
## [1] 1934.762
## [1] 2134.944

set.seed(1234)
result=kmeans(cost,3)
# mean cost of the clusters
result[2]

## $centers
##   cluster_data
## 1     2841.635
```

```
## 2    14143.293
## 3    28338.916

# DRG under each clusters
DRG_cluster=as.data.frame(result[1])
#write.csv(DRG_cluster, file="~/Downloads/DRG_cluster.csv")
cost_2=cost
cost_2$DRG=table$DRG
DRG_cluster_2=DRG_cluster
DRG_cluster_2$DRG=table$DRG
plot_data=merge(DRG_cluster_2,cost_2,by= "DRG",all=FALSE)
colnames(plot_data)=c("DRG","group","cost")

data=plot_data
data$group=as.character(data$group)

ggplot(data, aes(DRG, cost, color = group)) + geom_point()+
  labs(x = "DRG", y = "Cost($)", color = "Group\n")+ scale_color_manual(labels = c("low cost", "medium cost","
high cost"), values = c("#008080", "#ffcc5c","#d9534f"))+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```




```
colnames(data) <- c("Name", "group","cost")
low=data[data$group=="1",]
```

```
medium=data[data$group=="2",]
high=data[data$group=="3",]
low=merge(low,DRG,by="Name")
medium=merge(medium,DRG,by="Name")
high=merge(high,DRG,by="Name")
```

analysis of low cost

inp features - hospital share HHI: 0.1885219;hospital num HHI 0.1822806 top hospital is University of Vermont Medical Center (as of 2014) – admission type: Emergency is the most common - average pdays: 4.274434

```
inp_data=inp[,c(1,2,54,76)]

low_inp=merge(inp_data,low,by="DRG",all=FALSE)
low_inp=as.data.table(low_inp)
low_inp=low_inp[cost!=0]

hospital_share=low_inp[,.(cost=sum(cost)),keyby =.(hnum2)]
hospital_share$share=hospital_share$cost/(sum(hospital_share$cost))
hospital_share=hospital_share[order(-share)]
hospital_share$share_2=hospital_share$share*hospital_share$share
low_hhi_share=sum(hospital_share$share_2)

hospital_num=low_inp[,.(number=.N),keyby =.(hnum2)]
hospital_num$share=hospital_num$number/(sum(hospital_num$number))
hospital_num=hospital_num[order(-share)]
hospital_num$share_2=hospital_num$share*hospital_num$share
low_hhi_num=sum(hospital_num$share_2)


adtype=low_inp[,.(number=.N),keyby =.(ATYPE)]
adtype=adtype[order(-number)]

pdays=low_inp[,.(number=.N),keyby =.(pdays)]
pdays=pdays[order(-number)]
sum(pdays$number)

## [1] 37714

sum(pdays$pdays*pdays$number)/37714

## [1] 4.274434
```

revcode features

- **Top 5 expenditure of service: Room & Board (Semi-Private 2 beds) Laboratory - Clinical Diagnostic Emergency Room Pharmacy Psychiatric**

- **Top 5 common service: Pharmacy Drugs Require Specific ID: Drugs requiring detail coding Pharmacy: Other Drugs Require Specific ID: Self admin drugs (insulin admin in emergency-diabetes coma) Laboratory - Clinical Diagnostic**

```
rev_2=revcode[,c(5,6,7,8,10)]
inp_rev_2=merge(inp_new,rev_2,by="Uniq",all.x = FALSE)
inp_rev_2=inp_rev_2[,c(2:5)]
low_rev=merge(inp_rev_2,low,by="DRG",all=FALSE)
low_rev=low_rev[cost!=0]

REVCODE_share=low_rev[,.(total=sum(REVCHRGS)),keyby =.(REVCODE)]
REVCODE_share=REVCODE_share[order(-total)]

REVCODE_num=low_rev[,.(number=sum(REVUNITS)),keyby =.(REVCODE)]
REVCODE_num=REVCODE_num[order(-number)]
```

analysis of medium cost

inp features - hospital share HHI: 0.2590445;hospital num HHI 0.253887 top hospital is University of Vermont Medical Center (as of 2014) – admission type: Elective is the most common - average pdays: 4.673648

```
inp_data=inp[,c(1,2,54,76)]

medium_inp=merge(inp_data,medium,by="DRG",all=FALSE)
medium_inp=as.data.table(medium_inp)
medium_inp=medium_inp[cost!=0]

hospital_share=medium_inp[,.(cost=sum(cost)),keyby =.(hnum2)]
hospital_share$share=hospital_share$cost/(sum(hospital_share$cost))
hospital_share=hospital_share[order(-share)]
hospital_share$share_2=hospital_share$share*hospital_share$share
medium_hhi_share=sum(hospital_share$share_2)

hospital_num=medium_inp[,.(number=.N),keyby =.(hnum2)]
hospital_num$share=hospital_num$number/(sum(hospital_num$number))
hospital_num=hospital_num[order(-share)]
hospital_num$share_2=hospital_num$share*hospital_num$share
medium_hhi_num=sum(hospital_num$share_2)

adtype=medium_inp[,.(number=.N),keyby =.(ATYPE)]
adtype=adtype[order(-number)]

pdays=medium_inp[,.(number=.N),keyby =.(pdays)]
pdays=pdays[order(-number)]
sum(pdays$number)
```

```
## [1] 10504

sum(pdays$pdays*pdays$number)/10504

## [1] 4.673648
```

revcode features

- **Top 5 expenditure of service: Operating Room Services Medical/Surgical Supplies: Other implants Room & Board (Semi-Private 2 beds) Pharmacy Anesthesia**
- **Top 5 common service: Pharmacy Anesthesia Operating Room Services Drugs Require Specific ID: Drugs requiring detail coding Drugs Require Specific ID: Self admin drugs (insulin admin in emergency-diabetes coma)**

```
rev_2=revcode[,c(5,6,7,8,10)]
inp_rev_2=merge(inp_new,rev_2,by="Uniq",all.x = FALSE)
inp_rev_2=inp_rev_2[,c(2:5)]
medium_rev=merge(inp_rev_2,medium,by="DRG",all=FALSE)
medium_rev=medium_rev[cost!=0]

REVCODE_share=medium_rev[,.(total=sum(REVCHRGS)),keyby =.(REVCODE)]
REVCODE_share=REVCODE_share[order(-total)]

REVCODE_num=medium_rev[,.(number=sum(REVUNITS)),keyby =.(REVCODE)]
REVCODE_num=REVCODE_num[order(-number)]
```

analysis of high cost

```
inp features - hospital share HHI: 0.7596885; hospital num HHI 0.7545897 top hospital is University of
Vermont Medical Center (as of 2014) – admission type: Elective is the most common - average pdays: 9.486261

inp_data=inp[,c(1,2,54,76)]

high_inp=merge(inp_data,high,by="DRG",all=FALSE)
high_inp=as.data.table(high_inp)
high_inp=high_inp[cost!=0]

hospital_share=high_inp[,.(cost=sum(cost)),keyby =.(hnum2)]
hospital_share$share=hospital_share$cost/(sum(hospital_share$cost))
hospital_share=hospital_share[order(-share)]
hospital_share$share_2=hospital_share$share*hospital_share$share
high_hhi_share=sum(hospital_share$share_2)

hospital_num=high_inp[,.(number=.N),keyby =.(hnum2)]
hospital_num$share=hospital_num$number/(sum(hospital_num$number))
hospital_num=hospital_num[order(-share)]
```

```
hospital_num$share_2=hospital_num$share*hospital_num$share
high_hhi_num=sum(hospital_num$share_2)

adtype=high_inp[,.(number=.N),keyby =.(ATYPE)]
adtype=adtype[order(-number)]

pdays=high_inp[,.(number=.N),keyby =.(pdays)]
pdays=pdays[order(-number)]
sum(pdays$number)

## [1] 1201

sum(pdays$pdays*pdays$number)/1201

## [1] 9.486261
```

revcode features

- Top 5 expenditure of service: Operating Room Services Room & Board (Semi-Private 2 beds) Medical/Surgical Supplies: Other implants Intensive care Pharmacy
- Top 5 common service: Pharmacy Drugs Require Specific ID: Drugs requiring detail coding Operating Room Services Anesthesia Drugs Require Specific ID: Self admin drugs (insulin admin in emergency-diabetes coma)

```
rev_2=revcode[,c(5,6,7,8,10)]
inp_rev_2=merge(inp_new,rev_2,by="Uniq",all.x = FALSE)
inp_rev_2=inp_rev_2[,c(2:5)]
high_rev=merge(inp_rev_2,high,by="DRG",all=FALSE)
high_rev=high_rev[cost!=0]

REVCODE_share=high_rev[,.(total=sum(REVCHRGS)),keyby =.(REVCODE)]
REVCODE_share=REVCODE_share[order(-total)]

REVCODE_num=high_rev[,.(number=sum(REVUNITS)),keyby =.(REVCODE)]
REVCODE_num=REVCODE_num[order(-number)]

drg_cluster_low=as.data.frame(low_inp$Name)
colnames(drg_cluster_low)=c("name")
drg_cluster_low$name=as.character(factor(drg_cluster_low$name))

List <- strsplit(drg_cluster_low$name, " ")
List=data.frame(Words=unlist(List))
List=as.data.table(List)
List=List[,.(number=.N),keyby=.(Words)]
List=List[order(-number)]
List=List[-c(1,2,3,5,11,12,13,14,15,16,19,20,21)]
List=List[-c(13:16)]
List=List[-c(36)]
```

```
List=List[-c(3)]

#cloud_low=List %>%with(wordcloud(Words, number, max.words = 80,colors=brewer.pal(8, "Dark2")))

drg_cluster_medium=as.data.frame(medium_inp$Name)
colnames(drg_cluster_medium)=c("name")
drg_cluster_medium$name=as.character(factor(drg_cluster_medium$name))

List <- strsplit(drg_cluster_medium$name, " ")
List=data.frame(Words=unlist(List))
List=as.data.table(List)
List=List[,.(number=.N),keyby=.(Words)]
List=List[order(-number)]

List=List[-c(11,15)]



#cloud_low=List %>%with(wordcloud(Words, number, max.words = 80,colors=brewer.pal(8, "Dark2")))

inp_data=inp[,c(1,2,4,6,54,76)]
low_gender=merge(inp_data,low,by="DRG",all=FALSE)[,5]
low_gender=low_gender[,.(number=.N),keyby=.(sex)]
low_gender=low_gender[-1,]
17616+23540

## [1] 41156

low_gender=low_gender[,.(percentage=number/41156),keyby=.(sex)]

inp_data=inp[,c(1,2,4,6,54,76)]
medium_gender=merge(inp_data,medium,by="DRG",all=FALSE)[,5]
medium_gender=medium_gender[,.(number=.N),keyby=.(sex)]
medium_gender

##    sex number
## 1:   1   4867
## 2:   2   5637

4867+5637

## [1] 10504

medium_gender=medium_gender[,.(percentage=number/10504),keyby=.(sex)]

inp_data=inp[,c(1,2,4,6,54,76)]
high_gender=merge(inp_data,high,by="DRG",all=FALSE)[,5]
high_gender=high_gender[,.(number=.N),keyby=.(sex)]
753+448

## [1] 1201
```

```
high_gender=high_gender[,.(percentage=number/1201),keyby=.(sex)]
```

```
48.87745
```

```
age=fread(file="~/Downloads/age.csv",col.names=c('intage','median'),select = c('AGEGRP','median'))
inp_data=inp[,c(1,2,4,6,54,76)]
row_number=merge(inp_data,low,by="DRG",all=FALSE)[,4]
low_age=merge(inp_data,low,by="DRG",all=FALSE)[,4]
low_age=merge(low_age,age,by="intage",all=FALSE)

low_age=low_age[,.(number=.N),keyby=.(median)]
average=sum(low_age$median*low_age$number)/nrow(row_number)
```

```
60.04446
```

```
age=fread(file="~/Downloads/age.csv",col.names=c('intage','median'),select = c('AGEGRP','median'))
inp_data=inp[,c(1,2,4,6,54,76)]
row_number=merge(inp_data,medium,by="DRG",all=FALSE)[,4]
medium_age=merge(inp_data,medium,by="DRG",all=FALSE)[,4]
medium_age=merge(medium_age,age,by="intage",all=FALSE)

medium_age=medium_age[,.(number=.N),keyby=.(median)]
average=sum(medium_age$median*medium_age$number)/nrow(row_number)
```

```
60.81848
```

```
age=fread(file="~/Downloads/age.csv",col.names=c('intage','median'),select = c('AGEGRP','median'))
inp_data=inp[,c(1,2,4,6,54,76)]
row_number=merge(inp_data,high,by="DRG",all=FALSE)[,4]
high_age=merge(inp_data,high,by="DRG",all=FALSE)[,4]
high_age=merge(high_age,age,by="intage",all=FALSE)

high_age=high_age[,.(number=.N),keyby=.(median)]
average=sum(high_age$median*high_age$number)/nrow(row_number)
```

```
 from pyecharts import WordCloud
import matplotlib as plt
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
np.set_printoptions(suppress=True)
x11=pd.read_csv("~/Desktop/medium_output.csv",sep=',')
x11=x11.drop(columns=['Unnamed: 0'])
x11=x11.drop_duplicates()
x1_frame=x11
#combine with drg
data_filter1_name1=pd.merge(data_filter1_name,x1_frame,how='right',left_on='MSDRG',right_on='DRG')
data_filter1_name1=data_filter1_name1.dropna(how='any')
from pyecharts import WordCloud
import matplotlib as plt
import matplotlib.pyplot as plt
```

```
#Split
word_split=data_filter1_name1['MSDRG_DESC'].str.split(' ')
word_split_pd=pd.DataFrame(word_split)
word_split_pd['MSDRG_DESC']=word_split_pd['MSDRG_DESC'].astype(str)
word=[]
for x in word_split_pd['MSDRG_DESC']:
   word.extend(str.split(x, sep=','))
word_after=pd.Series(word)
word_after=word_after.str.replace('[', '')
word_after=word_after.str.replace(']', '')
word_after=word_after.str.replace("'", '')
word_after=word_after.str.replace(" ", '')
cccc=word_after.value_counts()
dict_temp1= {'Word':cccc.index, 'Total_number':cccc.values}
dict_temp1=pd.DataFrame(dict_temp1)
dict_temp1.to_csv("~/Desktop/word_split_x11.csv")
dict_temp1_delet=pd.read_csv("~/Desktop/word_split_x11_delete.csv",sep=',')
temp_name1=dict_temp1_delet['Word']
temp_value1=dict_temp1_delet['Total_number']
wordcloud =WordCloud(width=1300, height=620)
wordcloud.add("", temp_name1, temp_value1, word_size_range=[20, 100])
wordcloud.render()
```