

Insurance Claims Data & Analytics



December 4, 2019

Brandeis university

# **Study of RA Disease**

We have identified two patient cohorts using diagnosis codes reported in the claim data. The first cohort consists of patients diagnosed with common chronic Rheumatoid Arthritis (RA), and the second cohort consists of patients diagnosed with other RA with systemic involvement. By examining data with regards to the two cohorts, we came to the following findings:

* The Most Common Types of the RA

**Table 1** Most Common Types of RA (Cohort I)

|  |  |
| --- | --- |
| **ICD-10 CODE** | **FREQUENCY** |
| M069 | 909 |
| M0579 | 17 |
| M059 | 8 |

**Table 2** Most Common Types of RA (Cohort II)

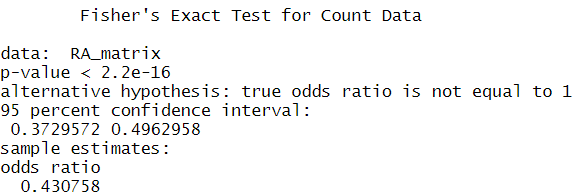
|  |  |
| --- | --- |
| **ICD-10 CODE** | **FREQUENCY** |
| M0510 | 21 |
| M0519 | 2 |
| M05671 | 2 |

* Gender Differences in RA

**Table 3** Cross Tabulation (RA Diagnosis by Sex)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RA Patients** | **Other Patients** | **Total** |
| **Male** | 266 | 168,136 | 168,402 |
| **Female** | 740 | 201,480 | 202,220 |
| **Total** | 1,006 | 369,616 | 370,622 |

We use Fisher’s Exact Test to test the gender difference in getting Rheumatoid Arthritis (RA). The null hypothesis is that there is no significant gender difference, and the alternative hypothesis is that there is significant gender difference. According to the results of Fishers’ Exact Test for the table (shown in **Figure 1**), the p-value is smaller than 2.2e-16 < 0.05, so we reject the null hypothesis at the 5% significance level that true odds ratio is equal to 1. Namely, we think that the gender preference difference in getting RA exists. The odds ratio 0.4308 means that, holding other factors constant, the odds of a female getting RA is 1.32% higher than that of a male. This result is intuitive because pretty many females like to wear dresses in cold winter. This makes them much more likely to get RA.



**Figure 1** Fisher’s Exact Test Results

* The Inter-Quartile Range of the Costs

**Table 4** Quartiles of Outpatient Charges

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **0%** | **25% (Q1)** | **50% (Q2)** | **75% (Q3)** | **100%** |
| 0 | 682 | 1522 | 3440 | 227312 |

The inter-quartile range is 2757.7 (Q1 - Q3).

* Study of Service Utilization

**Table 5** Top-5 Most Common Services for Treatment of the RA

|  |  |  |
| --- | --- | --- |
| **SERVICE** | **REVCODE** | **N** |
| Laboratory - Clinical Diagnostic | 300 | 3272 |
| Drugs Require Specific ID: Drugs requiring detail coding | 636 | 1230 |
| Pharmacy | 250 | 1168 |
| Emergency Room | 450 | 1065 |
| Radiology - Diagnostic | 320 | 377 |

# **Major Diagnostic Category (MDC) Concentration**

Among the two major diagnostic categories, we believe that MDC 14 – Pregnancy, Childbirth and Puerperium would be done more generally by most of the hospitals, while MDC 1 – Diseases and Disorders of the Brain and Nervous System would be highly concentrated among specialized high technology medical centers.

Intuitively, MDC 14 is a basic diagnostic category that should be performed by most hospitals, while MDC 1 is a diagnostic category related to complex or serious problems that require more professional treatment, which only a few “referral hospitals/centers” are able to deal with.

The analytics results of inpatient file state the same. The HHI calculated by patient counts is 4070.77 for MDC 1, and 2141.16 for MDC 14. The HHI calculated by total charges is 6368.35 for MDC 1, and 2453.17 for MDC 14. It is obvious that MDC 1 is more concentrated among few big players. Hospital number 5 - University of Vermont Medical Center holds the “lion share” of the market. The market share for MDC 1 by patient counts is 62.24%, by charges is 79.27%.

University of Vermont Medical Center serve a population of more than 1 million people in Vermont and northern New York and approximately 160,000 residents in Chittenden and Grand Isle counties. It is also designated a Level I Trauma Center[1] (a Level I Trauma Center is capable of providing total care for every aspect of injury – from prevention through rehabilitation).[2] University of Vermont Medical Center is one of the so called “referral hospitals/centers”, so it is reasonable that it holds the largest market share.

**Table 6** Market Share by Patient Counts and by Charges (MDC 1)

| **Hospital Number** | **Patient  Counts** | **Share  by Counts** | **Total  Charges/$** | **Share  by Charges** |
| --- | --- | --- | --- | --- |
| 1 | 111 | 3.39% | 1,287,275 | 1.11% |
| 2 | 62 | 1.89% | 1,069,212 | 0.92% |
| 3 | 60 | 1.83% | 1,369,112 | 1.18% |
| 4 | 55 | 1.68% | 448,451 | 0.39% |
| 5 | 2,039 | 62.24% | 92,200,270 | 79.27% |
| 6 | 204 | 6.23% | 3,813,673 | 3.28% |
| 8 | 335 | 10.23% | 9,202,322 | 7.91% |
| 9 | 41 | 1.25% | 715,207 | 0.61% |
| 10 | 85 | 2.59% | 1,604,507 | 1.38% |
| 11 | 15 | 0.46% | 220,861 | 0.19% |
| 12 | 62 | 1.89% | 1,015,393 | 0.87% |
| 14 | 15 | 0.46% | 187,743 | 0.16% |
| 15 | 57 | 1.74% | 955,792 | 0.82% |
| 16 | 135 | 4.12% | 2,219,602 | 1.91% |

**Table 7** Market Share by Patient Counts and by Charges (MDC 14)

| **Hospital  Number** | **Patient Counts** | **Share by Counts** | **Total Charges** | **Share by Charges** |
| --- | --- | --- | --- | --- |
| 1 | 378 | 6.66% | 2384525 | 3.61% |
| 2 | 206 | 3.63% | 2528758 | 3.82% |
| 3 | 226 | 3.98% | 2845267 | 4.30% |
| 4 | 206 | 3.63% | 1512194 | 2.29% |
| 5 | 2416 | 42.59% | 30773062 | 46.53% |
| 6 | 326 | 5.75% | 4049609 | 6.12% |
| 8 | 361 | 6.36% | 4171437 | 6.31% |
| 9 | 379 | 6.68% | 4824498 | 7.29% |
| 10 | 205 | 3.61% | 2947251 | 4.46% |
| 12 | 193 | 3.40% | 2144511 | 3.24% |
| 15 | 306 | 5.39% | 3053785 | 4.62% |
| 16 | 471 | 8.30% | 4907477 | 7.42% |

# **Cluster Analysis**

## 1. Background

In this project, we conduct a simple cluster analysis. Since clustering is the cornerstone of AI, computer returned the result of clustering based on its calculation. However, we don’t know the logic behind the result, since there is no science to tell us how to understand the properties of the clusters that mathematical models creating for us. Therefore, it’s our goal to try to interpret the clusters by looking into clusters of certain cost categories of the inpatient hospital DRGs (**Operating Room cost** and **Anesthesia cost**), to make sense of them using our own domain knowledge once they are formed by the machine.

## 2. Data Manipulation

* Filter the important DRGs between 20 and 977 from VTINP16, exclude REVCHRGS <100 in Revenue file, then merge two tables using the UNIQ variable to get the table which include all information.
* Sum all the charges by UNIQ, DRG, PCCR; cross tabulate the new table with DRGs (in the row) and the mean value of the PCCRs, as cell values.
* Plug in the PCCR and DRG real names into the table, create a new cost category as PCCR\_OR\_and\_Anesth\_Costs.
* Drop NAs and turn all those empty cells to zero dollars.
* Standardize variables for clustering.
* In the end we had the final table for clustering which only includes the DRG and PCCR\_OR\_and\_Anesth\_Costs.

The table below presents a grasp of the clustering data.

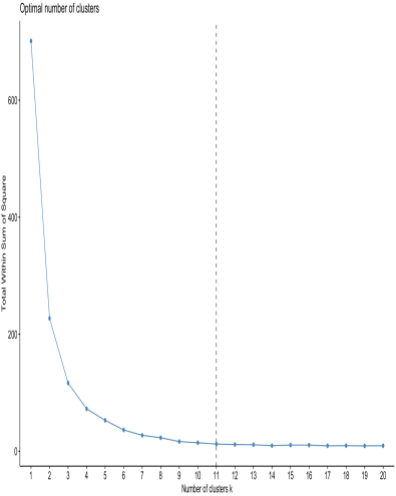
**Table 8** Head of the Clustering Data

| **DRG** | **PCCR\_OR\_and\_Anesth\_Costs** |
| --- | --- |
| 20 | 21805.87 |
| 21 | 15172.70 |
| 22 | 9857.05 |
| 23 | 17395.65 |
| 24 | 11151.29 |
| 25 | 21856.54 |
| 26 | 20890.97 |
| 27 | 16731.32 |
| 28 | 24159.53 |

## 3. Clustering

* **Comparing F-statistics**

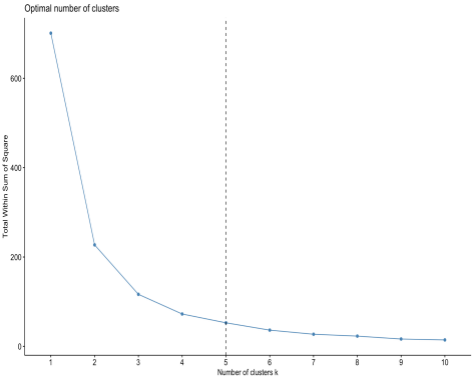
For clustering, we used k-means. Firstly, we tried to find the best k based on Calinski-Harabasz F-statistics. Our result shows that when k =11, the F-statistic is the highest, which is the best. Based on this, we reached the conclusion that, if we only use cost to do clustering, k =11 is the best. However, it requires that we only need to get the F-statistics when k = 2,3,4,5. In this case, we find that k= 5 has the highest F-statistic and is the best choice.



**Figure 2** Optimal Number of Clusters (I)

**Table 9** Calinski-Harabasz F-statistics (I)

|  |  |  |  |
| --- | --- | --- | --- |
| **Calinski-Harabasz F-statistics** | | | |
| 2 | 1460.13 | 8 | 2761.09 |
| 3 | 1753.53 | 9 | 3599.05 |
| 4 | 2018.13 | 10 | 3653.16 |
| 5 | 2141.12 | 11 | 3912.98 |
| 6 | 2545.71 | 12 | 3788.67 |
| 7 | 2872.02 | 13 | 3588.4 |



**Figure 3** Optimal Number of Clusters (II)

**Table 10** Calinski-Harabasz F-statistics (II)

|  |  |
| --- | --- |
| **Calinski-Harabasz F-statistics** | |
| 2 | 1460.133 |
| 3 | 1753.525 |
| 4 | 2018.129 |
| 5 | 2219.356 |

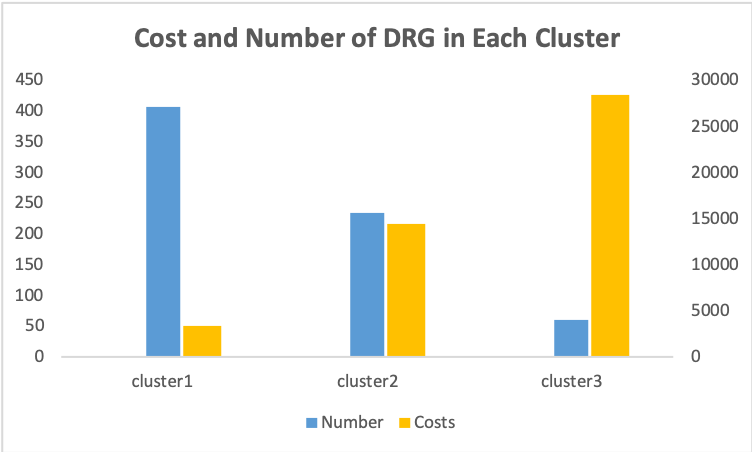
* **Clustering k=3**

We did our clustering when k = 3. The three clusters and costs are shown in **Figure 4** and **Figure 5**.

**Figure 4** the three Clusters of DRGsA screenshot of a cell phone

Description automatically generated

**Figure 5** Cost and Number of DRG in Each Cluster



From the above pictures, we can simply see that cluster1 is the lowest center cost DRG group with highest amount of number, followed by cluster2 with the second center cost DRG group and second large amount of number, Cluster3 is the highest cost DRG group with lowest amount of number. This makes sense because in general, there are more less severe diseases, and most of them are kind of cheap, while the amount of severe diseases should be much less but much more expensive.

## 4. Analysis Result

After we get the datasets of three clusters, we analyze each cluster respectively from six aspects, including diagnosis description, treatment type, patients’ demographics (age and sex), length of stay in hospital, admission type and discharge status. In order to analyze the similarities within each cluster, we merge the three DRG clusters with the inpatient file to get three inpatient claim clusters.

* **Diagnosis Description**

Intuitively, low cost diseases are more common and easy to cure, and high cost diseases are more rare and complex cases. The clustering results state similar.

**Table 11** Count of Unique DRG Code in each Cluster

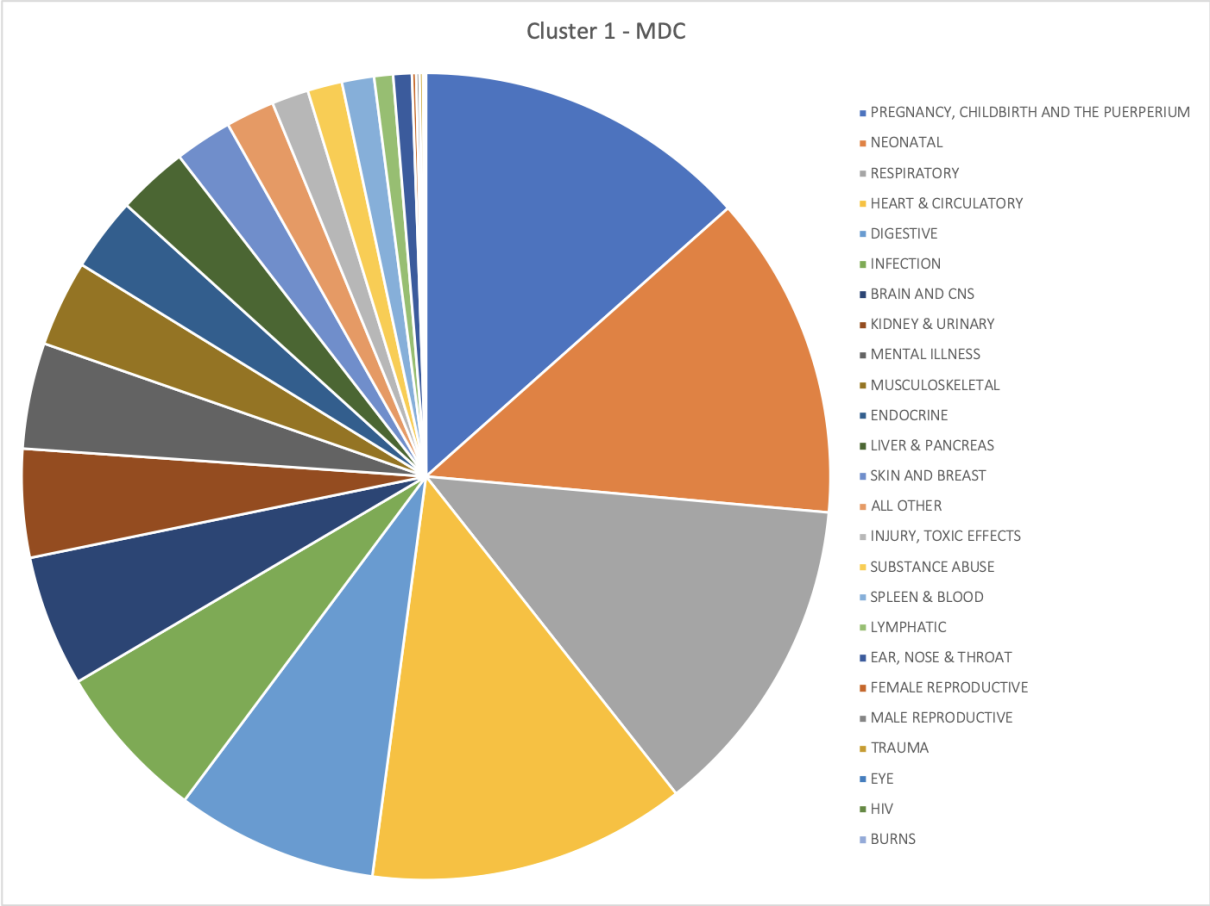
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| DRG count | 407 | 234 | 61 |

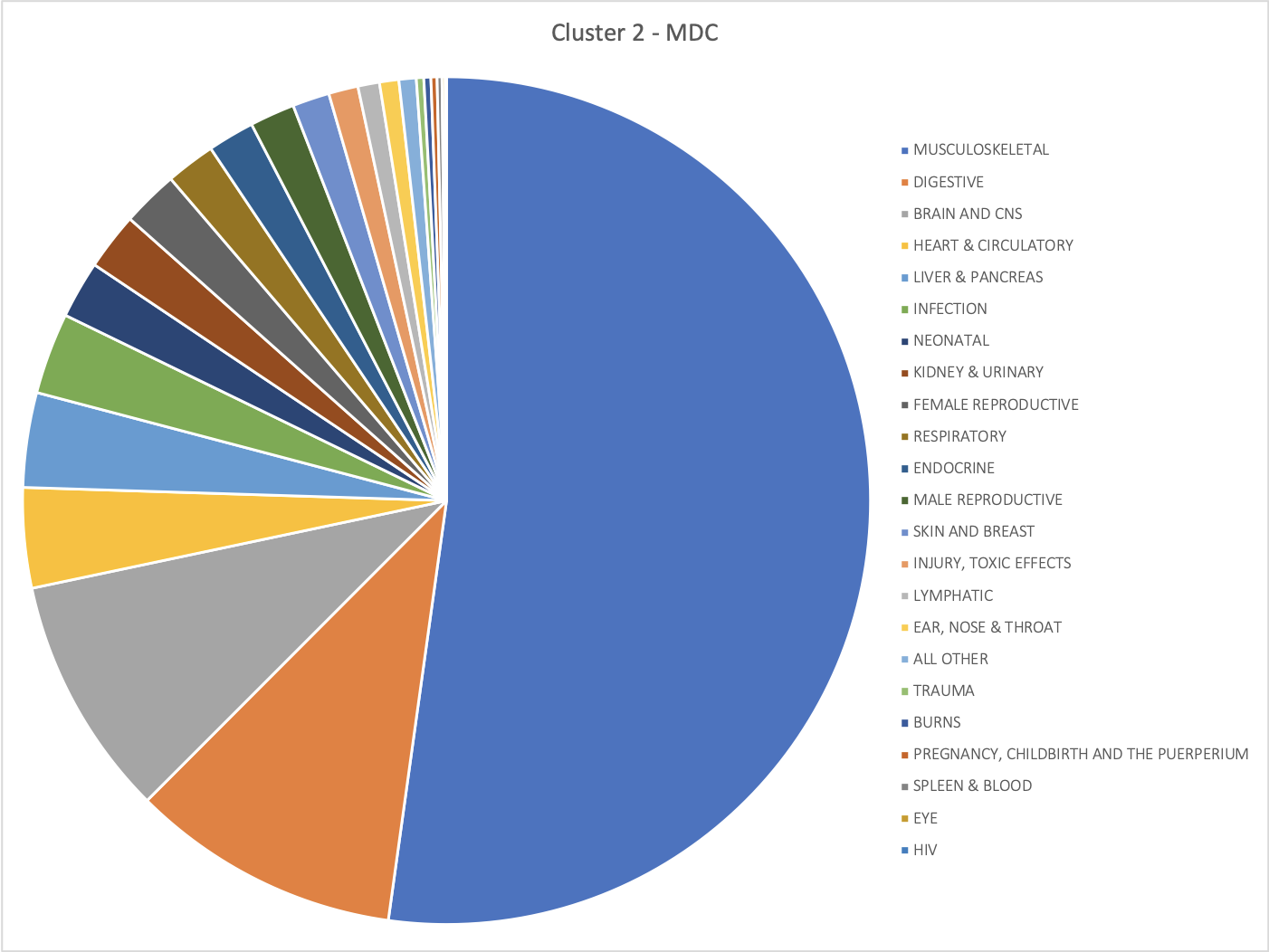
In order to get the diagnosis type in each cluster, we calculate the frequency of the DRG and the Major Diagnostic Categories (MDC) in each cluster. The top five most frequent DRG in each cluster and the pie charts of the percentage of MDC in each cluster are listed below.

**Table 12** Top Five of the Most Frequent DRG type in Each Cluster

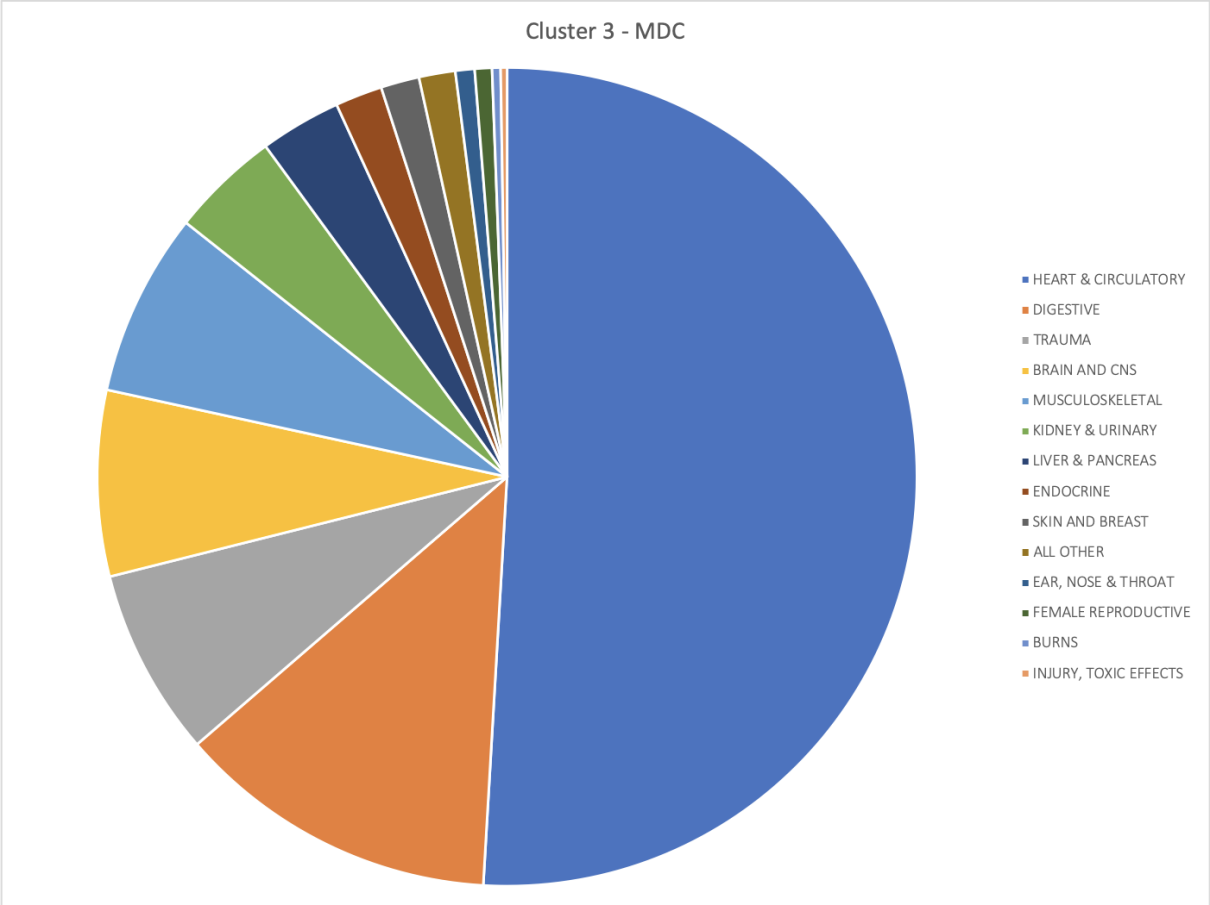
| **Cluster** | **Order** | **DRG Code** | **Count** | **Pct** | **DRG Description** |
| --- | --- | --- | --- | --- | --- |
| 1 | 1 | 775 | 3026 | 7.24% | KIDNEY TRANSPLANT |
| 2 | 795 | 2988 | 7.15% | OTHER KIDNEY & URINARY TRACT O.R. PROCEDURES |
| 3 | 794 | 1625 | 3.89% | URETHRAL PROCEDURES, AGE 0-17 |
| 4 | 871 | 1512 | 3.62% | UTERINE & ADNEXA PROC FOR NON-MALIGNANCY W/O CC |
| 5 | 885 | 1143 | 2.74% | INFECTIONS, FEMALE REPRODUCTIVE SYSTEM |
| 2 | 1 | 470 | 2354 | 22.79% | BRONCHITIS & ASTHMA AGE >17 W/O CC |
| 2 | 65 | 446 | 4.32% | ANAL & STOMAL PROCEDURES W/O CC |
| 3 | 460 | 389 | 3.77% | SIMPLE PNEUMONIA & PLEURISY AGE >17 WITH CC |
| 4 | 481 | 311 | 3.01% | CARDIAC VALVE & OTH MAJ CARDIOTHORACIC PROC W/O CARD CATH |
| 5 | 330 | 306 | 2.96% | EXTRAOCULAR PROCEDURES EXCEPT ORBIT AGE >17 |
| 3 | 1 | 329 | 115 | 9.70% | SEIZURE & HEADACHE AGE 0-17 |
| 2 | 220 | 84 | 7.08% | SPINAL DISORDERS & INJURIES |
| 3 | 253 | 78 | 6.58% | HYPERTENSIVE ENCEPHALOPATHY |
| 4 | 236 | 75 | 6.32% | CRANIAL & PERIPHERAL NERVE DISORDERS W/0 CC |
| 5 | 25 | 68 | 5.73% | NERVOUS SYSTEM INFECTION EXCEPT VIRAL MENINGITIS |

**Figure 6** Pie chart of MDC in Cluster 1



**Figure 7** Pie chart of MDC in Cluster 2

**Figure 8** Pie chart of MDC in Cluster 3



The diagnosis of cluster 1 are more spread, they are common diseases which won’t cause death in a short amount of time. According to the DRG frequency data, we can conclude that most cases are related to the urinary system problem (kidney transplant, other kidney and urinary tract or procedure, urethral procedure, otitis media), and non-malignant female reproductive system issues. From the MDC frequency data, we can see that most frequent cases are related to childbirth, afterbirth care and respiratory system problems. Even though the DRG data is very spread - the top five diagnosis related groups only take up less than 25% of all the data, it’s not representative enough. But we can still get the sense that these diseases are very likely or even inevitable to happen in daily life, and they have great probability to be cured.

The diagnosis of cluster are much concentrated. The DRG data suggests that the most frequent cases in cluster are related to respiratory system problems, such as bronchitis, asthma and simple pneumonia. In addition to that, digestive system problems also account for a high proportion. The MDC data states that the musculoskeletal issues are most frequent. The diseases in cluster 2 are more serious and urgent than cluster 1, but they are still very likely to be cured, however require more complex treatment procedure and longer time.

The diagnosis of cluster 3 are also quite concentrated. According to both the DRG and MDC datasets, the most frequent diseases are related to heart and circulatory system, such as seizure and headache. Besides, there are also a great amount of problems related to brain, nervous system problems and traumatic issues. These diseases are often urgent and may cause sudden death. From cluster 1 to cluster 3, the disease types become more rare, more serious and less likely to be cured.

* **Treatment Type**

As for the treatment for these three clusters, top five most frequent treatment of cluster1 are Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach which takes up 4.70% of the case, Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening which also takes up 4.70%, Respiratory Ventilation, 24-96 Consecutive Hours, Introduction of Nutritional Substance into Upper GI, Via Natural or Artificial Opening, Drainage of Peritoneal Cavity, Percutaneous Approach takes up 4.14% respectively.

**Table 13** Top 5 Frequent Treatment of Cluster 1

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 02HV33Z | 4.70% | Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach |
| 2 | 0BH17EZ | 4.70% | Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening |
| 3 | 5A1945Z | 4.14% | Respiratory Ventilation, 24-96 Consecutive Hours |
| 4 | 3E0G76Z | 4.14% | Introduction of Nutritional Substance into Upper GI, Via Natural or Artificial Opening |
| 5 | 0W9G3ZZ | 4.14% | Drainage of Peritoneal Cavity, Percutaneous Approach |
|  |  |  |  |

Top five most frequent treatment of cluster2 are Excision of Left Lower Leg Subcutaneous Tissue and Fascia, Open Approach which takes up 5.22% of the case, Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach which takes up 4.85%, Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening taking up 4.85%, Monitoring of Central Nervous Electrical Activity, External Approach taking up 3.73%, Ultrasonography of Superior Vena Cava, Guidance taking up 3.73%.

**Table 14** Top 5 Frequent Treatment of Cluster 2

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0JBP0ZZ | 5.22% | Excision of Left Lower Leg Subcutaneous Tissue and Fascia, Open Approach |
| 2 | 02HV33Z | 4.85% | Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach |
| 3 | 0BH17EZ | 4.85% | Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening |
| 4 | 4A10X4Z | 3.73% | Monitoring of Central Nervous Electrical Activity, External Approach |
| 5 | B548ZZA | 3.73% | Ultrasonography of Superior Vena Cava, Guidance |

**Table 15**  Top 5 frequent treatment of cluster3

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0BH17EZ | 4.67% | Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening |
| 2 | 02HV33Z | 4.67% | Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach |
| 3 | 3E033GC | 4.21% | Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach |
| 4 | 5A1935Z | 3.74% | Respiratory Ventilation, Less than 24 Consecutive Hours |
| 5 | 3E0G76Z | 3.74% | Introduction of Nutritional Substance into Upper GI, Via Natural or Artificial Opening |

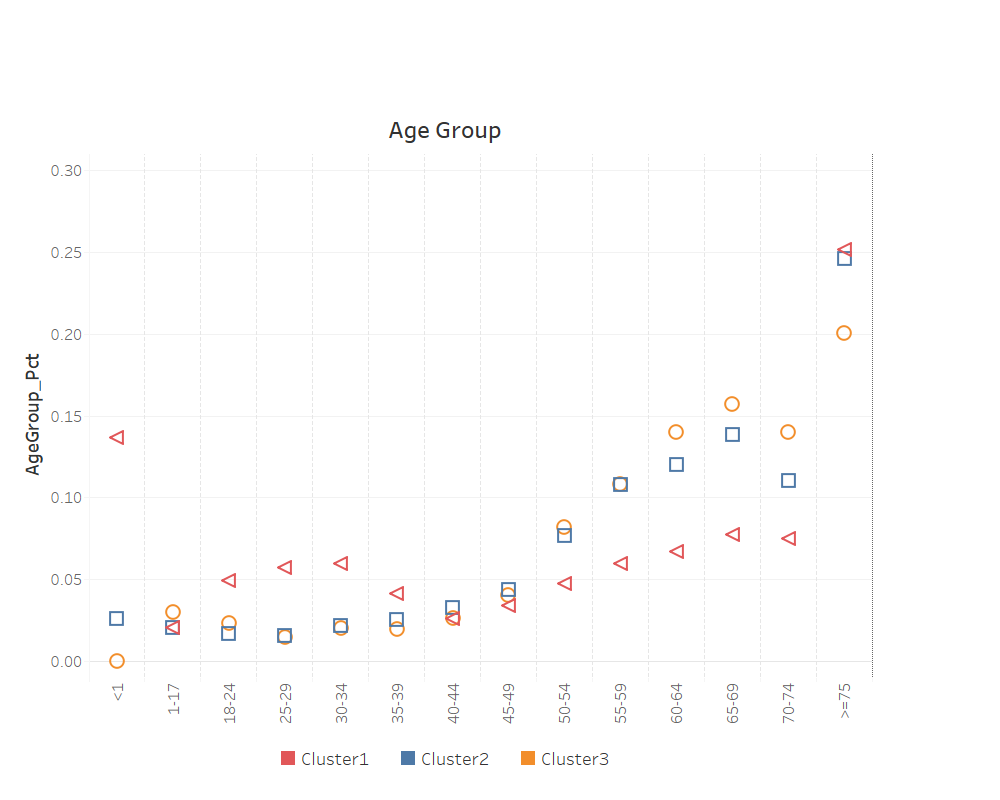
Top five most frequent treatment of cluster3 are Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening which takes up 4.67% of the case, Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach which takes up 4.67%, Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach taking up 4.21%, Respiratory Ventilation, Less than 24 Consecutive Hours taking up 3.74%, Introduction of Nutritional Substance into Upper GI, Via Natural or Artificial Opening taking up 3.74%.

We can see a lot of similarity between these three clusters. Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach，Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening both appeared in the top 5 most frequent treatment of all three clusters. So treatment may not be an effective factor to illustrate the similarity within a cluster and difference between them.

* **Patients’ Demographics (Age and Sex)**

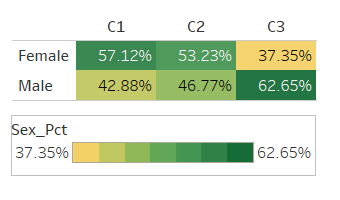
We consider patients’ demographics as another critical factor related to properties of the DRGs. Theoretically, older people are more likely to suffer from comorbidities or complications, so we think that a larger proportion of people of older age groups would be grouped into the DRGs Cluster with more complicated cases or higher costs. To find evidence in our data, we summarize the clusters by age group, calculating the percentage of each age group for each cluster. The results are consistent with our hypothesis. As shown in **Figure 9**, patients in DRGs in Cluster 2 and in Cluster 3 are more likely to be in older age groups as compared to those in Cluster 1, as the relative numbers of patients in DRGs in Cluster 2 and in Cluster 3 increase with age, while the patients with DRGs in Cluster 1 have relatively more patients ranging in age from 18 to 39. Besides, we see that there are relatively more patients ranging in age from 60 to 74 for Cluster 3 than for Cluster 2.

**Figure 9** Age Distribution of Patients in DRGs in the three Clusters



Another demographic factor we have examined is sex, taking into consideration the sex differences in patterns of disease due to both biological and social factors. For instance, generally speaking, men are more likely to have life-threatening diseases like stroke at young ages, while women are more likely to suffer from diseases such as anemia and arthritis and severe headaches. [3] Also, there exists gender gaps regarding level of personal resources such as income and education.

To find the sex distribution within and among the three clusters, we calculate the percentage of each sex for each cluster. From the results (shown in **Figure 10**), we notice that among the patients with DRGs in Cluster 3, males comprise a remarkably larger proportion, close to one third of the total. In comparison, the sex ratios for Cluster 1 and for Cluster 2 are less distorted, yet each of the cluster has a larger percentage of female patients than male patients. The results are consistent with our analysis regarding sex differences. Furthermore, the results relate to our previous analysis with respect to the frequencies of DRG in each cluster, as we have seen many diseases specific to females in Cluster 1 and more life-threatening diseases related to heart and circulatory system in Cluster 3.

**Figure 10** Sex Distribution of Patients in DRGs in the three Clusters 

* **Length of Stay in Hospital**

**Table 16** Summary Statistics of Length of Stay in Hospital

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Min** | **Max** | **Mean** |
| **Cluster 1** | 1 | 662 | 4.35 |
| **Cluster 2** | 1 | 213 | 4.51 |
| **Cluster 3** | 1 | 205 | 9.62 |

As we can see from **Table 16**, the average length of stay for Low Cost Cluster (Cluster 1) is 4.35 days, the average length of stay Middle Cost Cluster (Cluster 2) is 4.51 days and the High Cost Cluster (Cluster 3) is 9.62 days. This result is actually consistent with our assumption that the higher cost DRGs have longer length of stay in hospital.

For example, the most frequent MDC in cluster 1 is **PREGNANCY, CHILDBIRTH and THE PUERPERIUM.** Take Childbirth as an example.Even if it’s a very ideal plan to have a vaginal birth, sometimes a cesarean section is required to ensure the health of patients and their baby. Currently, a little more than 30 percent of deliveries in the United States are by c-section. Though every woman's situation is unique, the delivery of a baby via cesarean may take as little as 15 to 20 minutes, with a further 45 minutes needed to stitch up the uterus and abdominal incision. Usually you will stay in the hospital for a few extra days after a cesarean birth, so your doctors can make sure you are healthy and healing well. Most of childbirth surgeries are simple and easy. Patients usually don’t need to have follow-up surgeries or stay in hospital for a long time. Therefore, the cost of childbirth surgery is relatively low compared to other surgeries. The average total price charged for pregnancy and newborn care is about $30,000 for a vaginal delivery and $50,000 for a C-section.

When we look at MDC in cluster 3, the high cost cluster, we can find the most frequent MDC is **HEART & CIRCULATORY.** Take heart transplant surgery as an example. Average cost of a heart transplant surgery is $997,700. A heart transplant is a high-risk surgery that takes a patient about 6 months to reach full recovery. Just below the $1 million mark, this surgery takes a lot of preparation to ensure a healthy and best-match donor organ, as well as extensive recovery and follow-up procedures to ensure the patient’s prolonged health.

* **Admission Type**

**Table 17** Ratio of three Main Admission Types of Each Cluster

|  |  |  |  |
| --- | --- | --- | --- |
|  | **EM** | **urgent** | **elective** |
| **Low cost cluster** | 57% | 20% | 10% |
| **Middle cost cluster** | 35% | 9% | 52% |
| **High cost cluster** | 30% | 21% | 43% |

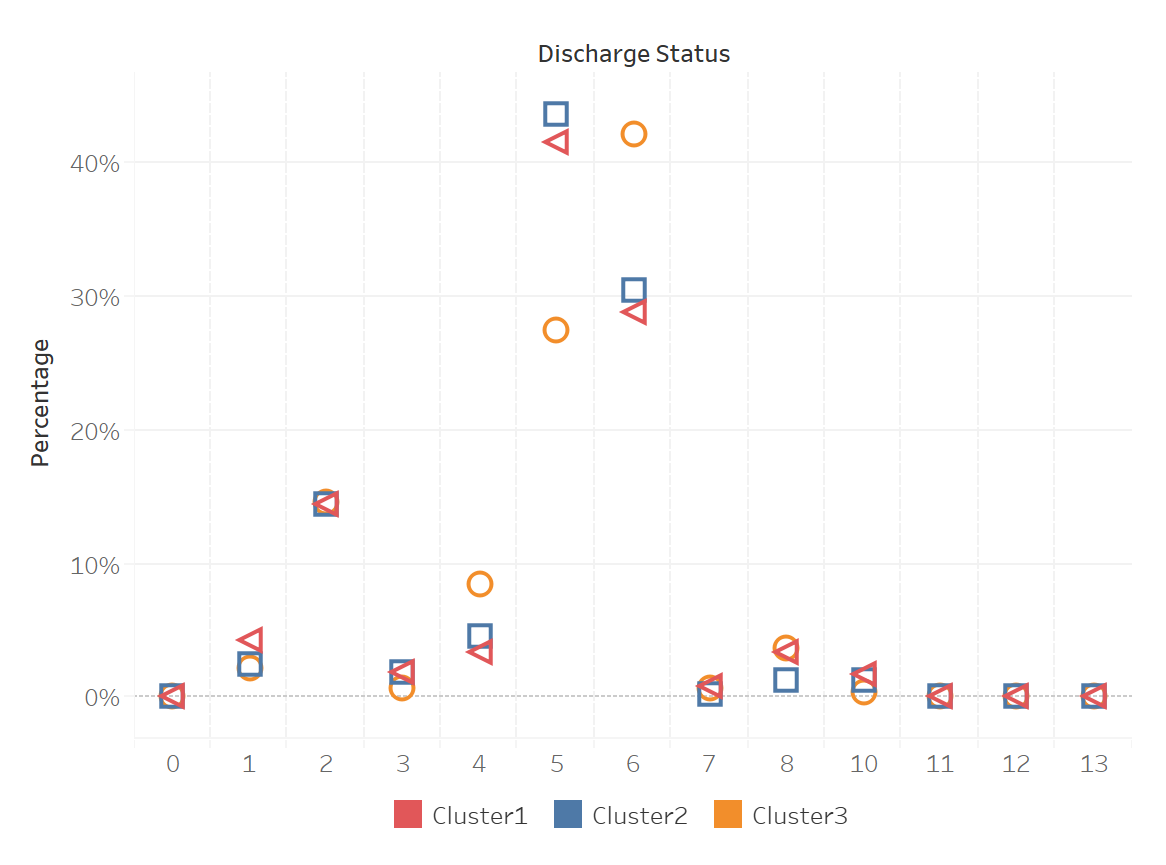
**Table 17** shows the ratio of three main admission types of each cluster. At first sight of the results, we might find that our result is kind of inconsistent with common sense.

However, after thinking, we realized this result makes sense. From the result, we consider middle cost group and high cost as a whole group to analyze because both of them have almost the same ratio of emergency and elective. When it comes to the interpretation, the low cost cluster includes some less serve diseases or injures which need to get emergency help. For example, there are car accidents, major bleeding, and elbow broken. Emergency are needed for those kinds of DRGs or they might be gonna die. However, those DRGs doesn’t need to stay so long in hospitals on average (about 4.3 days), because for example, for heavy blood bleeding, you only need to wrap the wound, have a rest, and then go home. Because it’s so emergent, and people don’t have chance choose, that’s why the elective ratio is low. However, for those DRGs in middle and high cost clusters, they are much more serve and complex which means they need more professional and complex treatment, and need to stay in hospital for a longer time(9.6/14.3 days) which is definitely more costly. People who have those kinds of diseases will choose more specialized services. That’s why the ratios of elective are relatively high compare with that of low cost cluster.

Another thing to notice is that in long period of time, elective type is much more expensive because more specialized doctors, complex surgeries and longer hospital stay days are needed. Those kind of things will be so expensive! On the contrary, for the emergency type, the doctors have comprehensive but not that specialized knowledge on various of diseases. Also, it will be less expensive because the average hospital stay days of this type of admission is much less.

* **Discharge Status**

We then studied the three clusters by the percentage distribution of Discharge Status. **Figure 11** is the Discharge Status distribution. To learn about the detailed description for each discharge , you could refer to **Table 18**. From **Figure 11**, we found that, in all three clusters, about 30%~40% of the patients have a discharge status of own/family care or home health. Among those patients, those treated with low/medium cost DRGs are more likely to go home and take an own/family care, while those treated with high cost DRGs are more likely to take home health. It might be because patients treated with high cost DRGs tend to pay more attention to their health conditions after being discharged from hospital, and they need to get some specific treatments by family doctors to prevent recurrence. In addition, we noticed a 5% higher percentage of transferring to another facility in those patients treated with high cost DRGs, compared with low/medium cost DRGs. It makes sense. Higher cost to some extent reflects the difficulty of treatment. If the high-cost treatment has little effect, the patients may transfer to another facility to seek for lower cost or better curative effect.



**Figure 11** Discharge Status Distribution of three Clusters

**Table 18** Discharge Status Description

|  |  |
| --- | --- |
| **Discharge Status** | **Discharge Status Description** |
| 1 | TO ANOTHER ACUTE HOSPITAL |
| 2 | TO A SKILLED NURSING FACILITY |
| 3 | TO AN INTERMEDIATE CARE FACILITY |
| 4 | TO ANOTHER FACILITY |
| 5 | HOME - OWN OR FAMILY CARE |
| 6 | HOME - HOME HEALTH |
| 7 | AGAINST MEDICAL ADVICE |
| 8 | DIED |
| 9 | HOME, CARE OF HOME IV PROVIDER |
| 10 | TO SWING BED |
| 11 | TO COURT OR LAW ENFORCEMENT |
| 12 | ADMITTED AS AN INPATIENT TO THIS HOSPITAL |
| 13 | STILL A PATIENT |
| 0 | MISSING |

## 5. Conclusion

Clustering DRGs by cost level, we noticed that Pregnancy, Childbirth and the Puerperium account for most in low cost cluster, while it is Musculoskeletal in medium cost cluster and Heart & Circulatory in cluster3. We found that patients treated with high cost DRGs are older, and males accounts for more. Diseases treated with high cost are rarer and more serious, so patients tend to have longer average length of stay in hospital and get home health care by family doctors after being discharged.  Besides, low cost DRGs have the highest emergency admission ratio with the lowest elective admission ratio, while middle and high cost clusters have relatively lower emergency admission ratio, but higher elective ratio. As for patient’s treatment, we see a lot of similarities between these three clusters, so treatment may not be an effective factor to illustrate the similarity within a cluster and difference between them.

Based on the above analysis, we think that it makes more sense to cluster the DRGs by diagnosis, demographics, length of hospital stay, admission type, and discharge status, instead of clustering only by cost.

# **Reference**

[1] [About the University of Vermont Medical Center - Burlington, Vt](https://www.uvmhealth.org/medcenter/Pages/About-UVM-Medical-Center.aspx)

[2] [Trauma Center Levels Explained - American Trauma Society](https://www.amtrauma.org/page/traumalevels)

[3] Gender Differences in Different Contexts

# **Appendix**

**# Question 1** **Study of RA Disease**

library(data.table)

library(dplyr)

setwd("C:/Users/ytlau/Desktop/Brandeis/193HS-256F-1 Healthcare Data Analytics and Data Mining/HW4")

# Load data

outp = fread("VTOUTP16.TXT")

RA\_1 = fread("chronic\_ra.csv")

RA\_2 = fread("other\_ra.csv")

# Take the ICD10\_Codes out as list

RA\_1\_codes = c(as.list(RA\_1[,1])[[1]])

RA\_2\_codes = c(as.list(RA\_2[,1])[[1]])

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

RA\_1\_cohort = outp[DX1 %in% RA\_1\_codes | DX2 %in% RA\_1\_codes | DX3 %in% RA\_1\_codes |

DX4 %in% RA\_1\_codes | DX5 %in% RA\_1\_codes | DX6 %in% RA\_1\_codes |

DX7 %in% RA\_1\_codes | DX8 %in% RA\_1\_codes | DX9 %in% RA\_1\_codes |

DX10 %in% RA\_1\_codes | DX11 %in% RA\_1\_codes | DX12 %in% RA\_1\_codes |

DX13 %in% RA\_1\_codes | DX14 %in% RA\_1\_codes | DX15 %in% RA\_1\_codes |

DX16 %in% RA\_1\_codes | DX17 %in% RA\_1\_codes | DX18 %in% RA\_1\_codes |

DX19 %in% RA\_1\_codes | DX20 %in% RA\_1\_codes]

RA\_2\_cohort = outp[DX1 %in% RA\_2\_codes | DX2 %in% RA\_2\_codes | DX3 %in% RA\_2\_codes |

DX4 %in% RA\_2\_codes | DX5 %in% RA\_2\_codes | DX6 %in% RA\_2\_codes |

DX7 %in% RA\_2\_codes | DX8 %in% RA\_2\_codes | DX9 %in% RA\_2\_codes |

DX10 %in% RA\_2\_codes | DX11 %in% RA\_2\_codes | DX12 %in% RA\_2\_codes |

DX13 %in% RA\_2\_codes | DX14 %in% RA\_2\_codes | DX15 %in% RA\_2\_codes |

DX16 %in% RA\_2\_codes | DX17 %in% RA\_2\_codes | DX18 %in% RA\_2\_codes |

DX19 %in% RA\_2\_codes | DX20 %in% RA\_2\_codes]

fwrite(RA\_1\_cohort,"RA\_1\_cohort.csv")

fwrite(RA\_2\_cohort,"RA\_2\_cohort.csv")

## Step 2: Identify the most common types of the RA

cohort\_1\_codes = c(RA\_1\_cohort[,DX1], RA\_1\_cohort[,DX2], RA\_1\_cohort[,DX3], RA\_1\_cohort[,DX4], RA\_1\_cohort[,DX5],

RA\_1\_cohort[,DX6], RA\_1\_cohort[,DX7], RA\_1\_cohort[,DX8], RA\_1\_cohort[,DX9], RA\_1\_cohort[,DX10],

RA\_1\_cohort[,DX11], RA\_1\_cohort[,DX12], RA\_1\_cohort[,DX13], RA\_1\_cohort[,DX14], RA\_1\_cohort[,DX15],

RA\_1\_cohort[,DX16], RA\_1\_cohort[,DX17], RA\_1\_cohort[,DX18], RA\_1\_cohort[,DX19], RA\_1\_cohort[,DX20])

RA\_1\_count = as.data.table(table(cohort\_1\_codes))[cohort\_1\_codes %in% RA\_1\_codes][

order(N,decreasing = T)]

fwrite(RA\_1\_count, "RA\_1\_count.csv")

cohort\_2\_codes = c(RA\_2\_cohort[,DX1], RA\_2\_cohort[,DX2], RA\_2\_cohort[,DX3], RA\_2\_cohort[,DX4], RA\_2\_cohort[,DX5],

RA\_2\_cohort[,DX6], RA\_2\_cohort[,DX7], RA\_2\_cohort[,DX8], RA\_2\_cohort[,DX9], RA\_2\_cohort[,DX10],

RA\_2\_cohort[,DX11], RA\_2\_cohort[,DX12], RA\_2\_cohort[,DX13], RA\_2\_cohort[,DX14], RA\_2\_cohort[,DX15],

RA\_2\_cohort[,DX16], RA\_2\_cohort[,DX17], RA\_2\_cohort[,DX18], RA\_2\_cohort[,DX19], RA\_2\_cohort[,DX20])

RA\_2\_count = as.data.table(table(cohort\_2\_codes))[cohort\_2\_codes %in% RA\_2\_codes][order(N,decreasing = T)]

fwrite(RA\_2\_count, "RA\_2\_count.csv")

## Step 3: Gender differences in RA

RA\_1\_M =RA\_1\_cohort[as.vector(RA\_1\_cohort[,"sex"] == "1"), ]

RA\_1\_F =RA\_1\_cohort[as.vector(RA\_1\_cohort[,"sex"] == "2"), ]

RA\_2\_M =RA\_2\_cohort[as.vector(RA\_2\_cohort[,"sex"] == "1"), ]

RA\_2\_F =RA\_2\_cohort[as.vector(RA\_2\_cohort[,"sex"] == "2"), ]

RA\_M\_count = count(RA\_1\_M) + count(RA\_2\_M)

RA\_F\_count = count(RA\_1\_F) + count(RA\_2\_F)

M = outp[as.vector(outp[,"sex"] == "1"), ]

F = outp[as.vector(outp[,"sex"] == "2"), ]

M\_count = count(M)

F\_count = count(F)

NRA\_M\_count = M\_count - RA\_M\_count

NRA\_F\_count = F\_count - RA\_F\_count

RA\_1\_matrix = matrix(unlist(c(count(RA\_1\_M), count(RA\_1\_F), NRA\_M\_count, NRA\_F\_count)), nrow=2)

fisher.test(RA\_1\_matrix)

RA\_2\_matrix = matrix(unlist(c(count(RA\_2\_M), count(RA\_2\_F), NRA\_M\_count, NRA\_F\_count)), nrow=2)

fisher.test(RA\_2\_matrix)

RA\_matrix = matrix(unlist(c(RA\_M\_count, RA\_F\_count, NRA\_M\_count, NRA\_F\_count)), nrow=2)

fisher.test(RA\_matrix)

## Step 4: Calculate the inter-quartile range of the costs

(quantile(outp$CHRGS))

(IQR\_chrgs = as.numeric(quantile(outp$CHRGS)[4] - quantile(outp$CHRGS)[2]))

## Step 5: Study of service utilization

revcode = fread("VTREVCODE16.TXT")

revcode\_dic = fread("revcode\_dic.csv")

rev\_service = merge(revcode, revcode\_dic, by = 'REVCODE')

revcode

RA\_1\_service = merge(RA\_1\_cohort, rev\_service, by = 'Uniq')

RA\_1\_topservice = as.data.table(table(RA\_1\_service$REVCODE\_DESC))[order(N,decreasing = T)]

RA\_2\_service = merge(RA\_2\_cohort, rev\_service, by = 'Uniq')

RA\_2\_topservice = as.data.table(table(RA\_2\_service$REVCODE\_DESC))[order(N,decreasing = T)]

RA\_service = merge(rbind(RA\_1\_cohort,RA\_2\_cohort), rev\_service, by = 'Uniq')

RA\_topservice = as.data.table(table(RA\_service$REVCODE\_DESC))[order(N,decreasing = T)]

RA\_topservice\_2 = unique(merge(RA\_service[,c("REVCODE","REVCODE\_DESC")], RA\_topservice, by.x = 'REVCODE\_DESC', by.y = 'V1'))[order(N,decreasing = T)]

fwrite(RA\_topservice\_2,"RA\_topservice.csv")

**# Question 2: Major Diagnostic Category (MDC) Concentration**

inp = fread("VTINP16\_upd.TXT")

# Select data of MDC = 1 and MDC = 14

MDC\_1 = inp[MDC==1,.(hnum2, CHRGS, MDC)]

MDC\_14 = inp[MDC==14, .(hnum2, CHRGS, MDC)]

# HHI calculation by patient counts

## MDC 1

share\_1 = as.data.table(table(MDC\_1$hnum2))

colnames(share\_1) = c("hnum2","count")

share\_1$market\_share = share\_1[,count/sum(count)]

(HHI\_1 = share\_1[, sum((market\_share\*100)^2)])

## MDC 14

share\_14 = as.data.table(table(MDC\_14$hnum2))

colnames(share\_14) = c("hnum2","count")

share\_14$market\_share = share\_14[,count/sum(count)]

(HHI\_14 = share\_14[, sum((market\_share\*100)^2)])

# --- MDC 1 is more concentrated

# HHI calculationby total charges

## MDC 1

char\_share\_1 = MDC\_1 [, .(charge\_by\_hospital = sum(CHRGS, na.rm = T)), by = hnum2]

char\_share\_1$market\_share = char\_share\_1[,charge\_by\_hospital/sum(charge\_by\_hospital)]

(char\_HHI\_1 = char\_share\_1[, sum((market\_share\*100)^2)])

## MDC 14

char\_share\_14 = MDC\_14[, .(charge\_by\_hospital = sum(CHRGS)), by = hnum2]

char\_share\_14$market\_share = char\_share\_14[,charge\_by\_hospital/sum(charge\_by\_hospital)]

(char\_HHI\_14 = char\_share\_14[, sum((market\_share\*100)^2)])

# --- MDC 1 is more concentrated

**# Question 3: Cluster Analysis**

library(data.table)

library(dbplyr)

setwd("/Users/tjmask/Desktop/Courses/health care/HW4")

## data manipulation

{

## loading data

# layout\_code = fread("FILE\_LAYOUT\_and\_CODES.csv")

revcode = fread("VTREVCODE16.csv")

inp = fread("VTINP16\_upd.csv")

head(revcode)

head(inp)

## filtering inp and revcode data

inp\_important = inp[DRG %in% c(20:977)]

revcode\_filter = revcode[!REVCHRGS<100]

## merging two tables

revcode\_inp = merge(inp\_important, revcode\_filter, by.x="UNIQ", by.y = "Uniq")

head(revcode\_inp)

## selecting useful columns from merged table

pccr\_drg = revcode\_inp[,.(total\_charge = sum(REVCHRGS)), by = .(UNIQ, DRG, PCCR)][order(PCCR,DRG,decreasing = FALSE)]

nrow(pccr\_drg)

setkey(pccr\_drg, DRG, PCCR)

## getting the avarage charge

avg\_charge = pccr\_drg[,.(avg\_charge = mean(total\_charge)), by = .(DRG, PCCR)]

avg\_charge= na.omit(avg\_charge)

## cross tabulation

tabulation = xtabs(avg\_charge ~ DRG + PCCR, avg\_charge)

## transfering the tabulation into data table

write.csv(tabulation, 'tabulation.csv')

table\_tabulation = fread('tabulation.csv', header= TRUE)

head(table\_tabulation)

## combining a new cost category

table\_tabulation$PCCR\_OR\_and\_Anesth\_Costs = table\_tabulation$`3700`+table\_tabulation$`4000`

## rename columns

pccr\_names = fread("pccr\_names.csv")

for (i in 1:length(colnames(table\_tabulation))){

for (j in 1:length(pccr\_names$PCCR)){

if (colnames(table\_tabulation)[i] == as.character(pccr\_names$PCCR[j])){

colnames(table\_tabulation)[i] = pccr\_names$PCCR\_NAME[j]

}

}

}

## save table table\_tabulation

write.csv(table\_tabulation, 'table\_tabulation.csv')

table\_tabulation = read.csv('table\_tabulation.csv', header= TRUE, row.names=1)

head(table\_tabulation)

nrow(table\_tabulation)

## getting the data for clustering

table\_cluster = table\_tabulation[c('PCCR\_OR\_and\_Anesth\_Costs')]

head(table\_cluster)

}

## clustering analysis

{

#install.packages("fpc")

#install.packages("factoextra")

library(factoextra)

library(cluster)

library(fpc)

library(tidyverse)

# To standarize the variables

table\_standarized = scale(table\_cluster)

head(table\_standarized)

## getting the Calinski-Harabasz F-statistics for k in (2,20)

f\_20 = c()

for (i in 2:13){

k.means.fit = kmeans(table\_cluster, i)

f\_20[length(f\_20)+1] = calinhara(table\_cluster, k.means.fit$cluster)

}

f\_20 = data.frame(f\_20)

rownames(f\_20) = c(2:13)

colnames(f\_20) = "F-statistic"

## drawing the picture of optimal number of clusters

fviz\_nbclust(table\_standarized, kmeans, k.max=20, method = "wss") +

geom\_vline(xintercept = 11, linetype = 2)

## getting the Calinski-Harabasz F-statistics for k in (2,5)

f\_5 = c()

for (i in 2:5){

k.means.fit = kmeans(table\_cluster, i)

f\_5[length(f\_5)+1] = calinhara(table\_cluster, k.means.fit$cluster)

}

f\_5 = data.frame(f\_5)

rownames(f\_5) = c(2:5)

colnames(f\_5) = "F-statistic"

## visualizing the SSE

fviz\_nbclust(table\_cluster, kmeans, method = "wss") +

geom\_vline(xintercept = 5, linetype = 2)

nrow(table\_cluster)

## clustering when k = 3

k.means.fit = kmeans(table\_cluster, 3) # k = 3

attributes(k.means.fit)

## getting the center of the 3 clusters

k.means.fit$centers

k.means.fit$cluster

## getting the size of each cluster

table(k.means.fit$cluster)

## data frame the clusters

df\_cluster = data.frame(k.means.fit$cluster)

## getting the specific rows of each cluster

cluster1 = rownames(df\_cluster)[k.means.fit$cluster==1]

cluster2 = rownames(df\_cluster)[k.means.fit$cluster==2]

cluster2

cluster3 = rownames(df\_cluster)[k.means.fit$cluster==3]

cluster3

## adding the detail description of each cluster

drg\_detail = fread("DRG\_detail.csv")

drg\_detail = as.data.frame(drg\_detail)

cluster11 = as.data.frame(cluster1)

cluster11$drg\_descp = drg\_detail[cluster11$cluster1,]$DRG\_DESC

cluster22 = as.data.frame(cluster2)

cluster22$drg\_descp = drg\_detail[cluster22$cluster2,]$DRG\_DESC

cluster33 = as.data.frame(cluster3)

cluster33$drg\_descp = drg\_detail[cluster33$cluster3,]$DRG\_DESC

write.csv(table\_cluster, "cluster\_cost.csv")

write.csv(cluster11, "cluster11.csv")

write.csv(cluster22, "cluster22.csv")

write.csv(cluster33, "cluster33.csv")

}

## visualization of the results

{

library(plotly)

## rearranging the data

cluster\_cost\_num = as.data.frame(k.means.fit$centers)

cluster\_cost\_num$number\_of\_drgs = c(407, 234,61)

cluster\_cost\_num$clusters = c("cluster1", "cluster2", "cluster3")

cluster\_cost\_num

write.csv(cluster\_cost\_num, 'cluster\_cost\_num.csv')

## plotting

p <- plot\_ly(data = cluster\_cost\_num) %>%

add\_trace(x = ~clusters, y = ~number\_of\_drgs, type = 'bar',

name="numbers",

marker = list(color = 'rgb(158,202,225)',

line = list(color = 'rgb(8,48,107)', width = 1.5))) %>%

add\_trace(x = ~clusters, y = ~PCCR\_OR\_and\_Anesth\_Costs,type = 'scatter',

name="costs", yaxis='y2',

marker = list(color = 'red',

line = list(color = 'red', width = 1.5))) %>%

layout(title = "numbers and cost of each cluster",

barmode = 'group',

xaxis = list(title = "clusters"),

yaxis = list(title = ""),

yaxis2=list(anchor='x', overlaying='y', side='right'))

p

library(plotly)

table\_clusters = read.csv("cluster\_cost.csv")

head(table\_clusters)

nrow(table\_clusters)

table\_clusters$clusters = 0

table\_clusters$clusters[which(table\_clusters$X %in% as.numeric(cluster1))] = "cluster1"

table\_clusters$clusters[which(table\_clusters$X %in% as.numeric(cluster2))] = "cluster2"

table\_clusters$clusters[which(table\_clusters$X %in% as.numeric(cluster3))] = "cluster3"

colnames(table\_clusters) =c("DRG","Costs","Clusters")

head(table\_clusters)

p2 <- plot\_ly(data = table\_clusters, x = ~DRG, y = ~Costs, color = ~Clusters)

p2

}

## further analysis

{

head(pccr\_drg)

length(unique(pccr\_drg$UNIQ))

length(unique(pccr\_drg$DRG))

head(revcode\_inp)

nrow(revcode\_inp)

## getting all the info for each cluster

which(colnames(revcode\_inp)=="DRG")

cluster1\_all = unique(revcode\_inp[DRG %in% as.numeric(cluster1)][,c(2, 10,31:50,76)])

cluster2\_all = unique(revcode\_inp[DRG %in% as.numeric(cluster2)][,c(2, 31:50,76)])

cluster3\_all = unique(revcode\_inp[DRG %in% as.numeric(cluster3)][,c(2, 31:50,76)])

## getting all the services info for each cluster

cluster1\_px = data.table(unique(gather(cluster1\_all[,3:22]))[2])

cluster2\_px = data.table(unique(gather(cluster2\_all[,3:22]))[2])

cluster3\_px = data.table(unique(gather(cluster3\_all[,3:22]))[2])

## find the top 30 services for each cluster

### top 30 services for cluster 1

top\_service\_cluster1 = cluster1\_px[,.(total\_num = .N), by =.(value)]

top\_service\_cluster1 = top\_service\_cluster1[order(top\_service\_cluster1[,2],

decreasing = T),][-1]

top30\_cluster1 = top\_service\_cluster1[1:30]

### top 30 services for cluster 2

top\_service\_cluster2 = cluster2\_px[,.(total\_num = .N), by =.(value)]

top\_service\_cluster2 = top\_service\_cluster2[order(top\_service\_cluster2[,2],

decreasing = T),][-1]

top30\_cluster2 = top\_service\_cluster2[1:30]

### top 30 services for cluster 3

top\_service\_cluster3 = cluster3\_px[,.(total\_num = .N), by =.(value)]

top\_service\_cluster3 = top\_service\_cluster3[order(top\_service\_cluster3[,2],

decreasing = T),][-1]

top30\_cluster3 = top\_service\_cluster3[1:30]

## export top 30 services of each cluster

write.csv(top30\_cluster1, "top30\_cluster1.csv")

write.csv(top30\_cluster2, "top30\_cluster2.csv")

write.csv(top30\_cluster3, "top30\_cluster3.csv")

## getting the hosipital frequency for each cluster

hp\_cluster1 =data.table(table(cluster1\_all$hnum2.x))

hp\_cluster1 = hp\_cluster1[order(hp\_cluster1[,2],decreasing=T),]

names(hp\_cluster1) = c("hnum","num\_treatment")

hp\_cluster2 =data.table(table(cluster2\_all$hnum2.x))

hp\_cluster2 = hp\_cluster2[order(hp\_cluster2[,2],decreasing=T),]

names(hp\_cluster2) = c("hnum","num\_treatment")

hp\_cluster3 =data.table(table(cluster3\_all$hnum2.x))

hp\_cluster3 = hp\_cluster3[order(hp\_cluster3[,2],decreasing=T),]

colnames(hp\_cluster3) = c("hnum","num\_treatment")

hp\_cluster1$hnum

hp\_cluster2$hnum

hp\_cluster3$hnum

## export hospital treatment frequency of each cluster

write.csv(hp\_cluster1, "hp\_cluster1.csv")

write.csv(hp\_cluster2, "hp\_cluster2.csv")

write.csv(hp\_cluster3, "hp\_cluster3.csv")

## getting drg frequency of each cluster

#cluster 1

freq\_cluster1 = inp\_important[inp\_important$DRG %in% as.numeric(cluster1)][,.(freq =.N), by = .(DRG)]

freq\_cluster1 = as.data.frame(freq\_cluster1)

freq\_cluster1 = freq\_cluster1[order(freq\_cluster1[,2], decreasing=T),]

head(freq\_cluster1,20)

#cluster 2

freq\_cluster2 = inp\_important[inp\_important$DRG %in% as.numeric(cluster2)][,.(freq =.N), by = .(DRG)]

freq\_cluster2 = as.data.frame(freq\_cluster2)

freq\_cluster2 = freq\_cluster2[order(freq\_cluster2[,2], decreasing=T),]

head(freq\_cluster2)

#cluster 3

freq\_cluster3 = inp\_important[inp\_important$DRG %in% as.numeric(cluster3)][,.(freq =.N), by = .(DRG)]

freq\_cluster3 = as.data.frame(freq\_cluster3)

freq\_cluster3 = freq\_cluster3[order(freq\_cluster3[,2], decreasing=T),]

head(freq\_cluster3)

## export drg frequency of each cluster

write.csv(freq\_cluster1, "freq\_cluster1.csv")

write.csv(freq\_cluster2, "freq\_cluster2.csv")

write.csv(freq\_cluster3, "freq\_cluster3.csv")

## average admission days of each cluster

# cluster 1

length(unique(pdays\_cluster1$DRG))

pdays\_cluster1 = inp\_important[inp\_important$DRG %in% freq\_cluster1$DRG][,.(DRG,pdays)]

pdays\_cluster1 = pdays\_cluster1[order(pdays\_cluster1[,2], decreasing = T)][5:nrow(pdays\_cluster1)]

pdays\_cluster1 = pdays\_cluster1[,.(sum(pdays)/.N)]

# cluster 2

pdays\_cluster2 = inp\_important[inp\_important$DRG %in% freq\_cluster2$DRG][,.(DRG,pdays)]

pdays\_cluster2 = pdays\_cluster2[order(pdays\_cluster2[,2], decreasing = T)]

pdays\_cluster2 = pdays\_cluster2[,.(sum(pdays)/.N)]

# cluster 3

pdays\_cluster3 = inp\_important[inp\_important$DRG %in% freq\_cluster3$DRG][,.(DRG,pdays)]

pdays\_cluster3 = pdays\_cluster3[order(pdays\_cluster3[,2], decreasing = T)]

pdays\_cluster3\_higher = pdays\_cluster3[pdays>=7]

pdays\_cluster3\_higher = pdays\_cluster3\_higher[,.(sum(pdays)/.N)]

pdays\_cluster3 = pdays\_cluster3[,.(sum(pdays)/.N)]

## admission type

# cluster 1

type\_cluster1 = inp\_important[inp\_important$DRG %in% freq\_cluster1$DRG][,.(DRG,ATYPE)]

type\_cluster1 = type\_cluster1[,.(type\_num=.N), by=.(ATYPE)]

type\_cluster1 = type\_cluster1[order(type\_cluster1[,2], decreasing = T)]

type\_cluster1\_ratio = type\_cluster1[,.(ATYPE, ratio = round((type\_num)/sum(type\_num),2))]

# cluster 2

type\_cluster2 = inp\_important[inp\_important$DRG %in% freq\_cluster2$DRG][,.(DRG,ATYPE)]

type\_cluster2 = type\_cluster2[,.(type\_num=.N), by=.(ATYPE)]

type\_cluster2 = type\_cluster2[order(type\_cluster2[,2], decreasing = T)]

type\_cluster2\_ratio = type\_cluster2[,.(ATYPE, ratio = round((type\_num)/sum(type\_num),2))]

# cluster 3

type\_cluster3 = inp\_important[inp\_important$DRG %in% freq\_cluster3$DRG[1:30]][,.(DRG,ATYPE)]

type\_cluster3 = type\_cluster3[,.(type\_num=.N), by=.(ATYPE)]

type\_cluster3 = type\_cluster3[order(type\_cluster3[,2], decreasing = T)]

type\_cluster3\_ratio = type\_cluster3[,.(ATYPE, ratio = round((type\_num)/sum(type\_num),2))]

}