

HW4: Insurance Claims Data & Analytics Part II

HW + your team slides for the final presentation are due Wednesday, Dec 4th, 12 PM (2 hours before the class starts).

Important note: while the HW includes all three questions, you will present only Q3 Cluster Analysis in class so your slides should contain only Q3 materials.

Context:

This homework lays out some analytics on the most important database in market for health, the insurance claim data. We use the publicly available All Payer Claim Data APCD you already used for previous HW. These databases consist of inpatient discharge data, outpatient procedures and services data, emergency department data, and revenue code file reporting on the charges associated with each of the service items rendered to the patient. Each data set according to the website includes:

- Case-specific diagnostic discharge data
- Some socio-demographic characteristics of the patient
- Medical reason for the admission
- Treatment and services provided to the patient
- Duration and status of the patient's stay in the hospital
- Full, undiscounted total and service-specific charges billed by the hospital

Question 1: Study of A Disease Cohort – data used: OutPatient + RevCode

Sometimes healthcare researchers need to focus on a narrowly defined cohort of patients with certain health conditions. For example study of cost and service utilization by patient with diabetes or cost drivers of treatment of patients with depression, etc. This kind of research agenda is very common and we want to exercise it here. In general, the research includes some basic steps: identify the cohort and major sub-cohorts; explore the cohort's demographics; study the service utilization and find patterns; study the costs; study the outcome of the treatment.

In this example we want to study the patients with **Rheumatoid Arthritis (RA)**. It is a chronic condition and many people suffer from the pain, discomfort, and the morbidities associated with it.

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

We identify the patient cohorts using diagnosis codes (DX codes) reported in the claim data. To save you time and effort I have searched the medical literature and have provided you with the ICD-10 DX codes of the two major clinical chapters of RA. However, in the real world this is a task you need to perform on your own perhaps with the help from a clinician or epidemiologist in your team. The ICD codes are listed in the Excel file named RA_ICD10_Codes.xlsx. One chapter is the common chronic RA and the other is other Rheumatoid Arthritis with systemic involvement (2 tabs in the Excel file). Since the RA patients are

mostly treated in the outpatient setting we need to search the OP file to identify the cohort. Note that the ICD-10 codes pertaining to RA can appear in any positions within the DX1-DX20 codes in the OP file. So your programming codes should be done so that for each outpatient encounter in the OP file, any of the DX1 to DX20 variables matching any of the RA ICD-10 codes in the Excel file is identified in the RA cohort for further analysis. If your code does not scan all of the DX1 to DX20 separately for a match against the RA ICD-10 codes you will have fewer patients in your cohort (under counting problem). You define two sub-cohorts separately, one cohort for the first tab in the Excel file (chronic RA) and the second cohort based on the Excel file tab “other RA with systemic involvement”.

Step 2: Identify the most common types of the RA

Each pt in your two sub-cohorts has at least one ICD-10 code by which the doctor diagnosed him/her as RA. Get the frequency of each of the ICD-10 codes and report the top-3 ICD-10 codes that are most prevalent in each sub-cohort.

Step 3: Gender differences in RA

Do you observe any pattern in RA prevalence (like one specific gender suffers more from the condition)? Is your observation statistically significant? (ugh, here we go again with the lovely Fisher’s Exact Test for a 2x2 X-tabulation ☹)

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

In statistics, quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are the 1st, 2nd, and 3rd quartiles also denoted by Q1, Q2, and Q3. The Q2 is what we know as median which divides the sample by half. Using the charge variable in the OP file (CHRGs) calculate the lower and upper values for the 4 quartiles and from there the inter-quartile range IQR.

Step 5: Study of service utilization

Link your two sub-cohort to the Revenue Code file using the UNIQ ID variable. List the top-5 most common services (as defined by the Revenue Codes) for treatment of the RA.



We all wish the APCD had also given us the Pharmacy Claims like other claims so we could mine that dataset as well to learn what type of medications (Rx) are prescribed for the RA patients and which doctor(s) tends to use more brand versus generic drugs.

But don’t worry in your real job you do get the pharmacy claims as well to investigate those exciting Rx related research and marketing questions. For now we need to table this piece.



Question 2: Which clinical chapter as defined by the Major Diagnostic Category MDC in inpatient care is more concentrated among few big players? – data used: Inp.

Not all hospitals are willing and able to perform surgical procedures and other inpatient medical care in all areas of medicine. Some hospitals are well equipped with technology and medical staff like academic medical centers. These so called “referral hospitals/centers” can take on any challenge in medicine and operate on virtually any patient no matter how complex the case is. In order to investigate this compare the two MDC 1 : Diseases and Disorders of the Brain and Nervous System and MDC 14 : Pregnancy, Childbirth And Puerperium. Before using the analytic part of your brain discuss among yourself as a team and make an informed guess as which MDC would be done more generally by most of the hospitals and which one tends to be highly concentrated among specialized high technology medical centers. Then turn to analytics using Inpatient file calculate the HHI index you used earlier in the Insurance market to investigate the concentration of the care for the two MDCs. Please repeat the HHI calculation by patient counts and by total charges in \$. The resulting HHIs by charges and by number of hospital admissions both should give you the answer. In the more concentrated MDC which hospital holds the “lion share” of the market? How much is the share that the lion gets (separately for admission counts and for dollars)? Google it and say a few words about the hospital. Now it’s time to check the results against your initial informed guess. Were you right in informed guessing before crunching the numbers? I bet 😊

Question 3: Clustering Costs – data used: Inp and RevCode

In this question we will conduct a simple cluster analysis. Clustering is the cornerstone of AI and for the most part the technique follows the same basic ideas as explained in this question. Here we look into clusters of certain cost categories of the inpatient hospital DRGs and try to interpret the clusters, to make sense of them using our own knowledge once they are formed by the machine. Understanding and making sense of the clusters is an art and could be mastered only by experience and of course a rich domain knowledge. As a matter of fact there is no science to tell us how to understand the properties of the clusters that mathematical models creating for us.

Inpatient hospitalizations are identified and priced and paid by DRGs. So each hospital admission has a DRG. You can see the list of DRGs in the Excel file “*FILE_LAYOUT_and_CODES*” available on LATTE. While the inpatient file provides you with a single summary dollar value for hospital charges, you can see all the details of those charges in the Revenue Code file *VTREVCODE16*. In the Revenue file there are more than one standard to classify the detailed services. The most common coding system is Revenue Code which is shown the variable RevCode in the file. There is an alternative coding system to group the revenue items known as Primary Cost Center or PCCR which classifies the revenues based on the department where the bill is originated. PCCRs are listed in the Excel file “*REVCODE_FILE_LAYOUT_and_CODES*”

Start from the Inpatient file VTINP16. Filter your hospital admissions to only important DRGs between 20 and 977. We put aside other DRGs for this analysis. Next, link your filtered Inpatient data to the Revenue Code file using the UNIQ variable. In your Revenue file exclude the low dollar value services (less than \$100) by dropping the REVCHRGs <100. Now sum all the charges by the PCCR category because some hospital admissions may have multiple charges submitted from the same PCCR department so you want to add them all. (In your SQL you can use something like: Select UNIQ, DRG, PCCR, Sum(REVCHRGs) Group By UNIQ, DRG, PCCR From My_IP_RevCode_Merged_Table).

Now that you have summed together all dollars for PCCR per hospital admission, you need to cross tabulate your list of selected DRGs (in the row) and the mean value of the PCCRs, as cell values. This cross-tabulation will yield approximately 687 rows (one per DRG) and 54 columns, one column per PCCR for all your PCCRs. (Your row numbers might differ by a few DRGs depending on your filtering of the data in the aforementioned steps. However, the number of DRGs in your row should not be substantially different from the 687. If you are higher or lower by more than a few then revisit your codes.) While your column headings are named by the PCCR codes you want to plug in the PCCR real name to your Revenue File and use the PCCR names instead, otherwise your cross-tabulation column headings are a bunch of PCCR numbers and that column naming would not serve you well in moving to cluster analysis. So you want your column heading to read “Physical Therapy” instead of PCCR 5000 and “Ultra Sound” instead of PCCR 3630. Also make sure you plug in the DRG names to your cross-tabulated file so you will have the DRG number and the name both as your row headings.

For this exercise, we want to focus on the two specific PCCR revenue centers (these are your column headings):

PCCR 3700 Operating Room

PCCR 4000 Anesthesiology

Create a new cost category by combining the **PCCR 3700 Operating Room + PCCR 4000 Anesthesiology**. Name this new cost category as **PCCR_OR_and_Anesth_Costs**. So far your x-tab should have DRG names as the row headings, the name of all 54 PCCRs +1 combined PCCR as the column heading, and the average costs (\$) of each PCCR category as the cell values.

You will soon notice that many cells in your x-tab are empty. This is expected, because for example pregnancy DRG has no Nuclear Medicine-Diagnostic service so that cell will be empty. In order to avoid any problem with your cluster analysis software you want to turn all those empty cells to zero dollars.

Cluster Analysis:

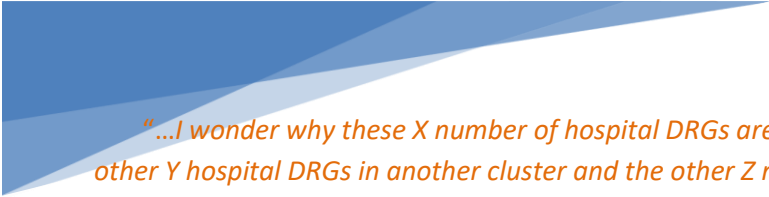
With your x-tab cleaned and formatted neatly and with no empty cells, you are ready to move to the cluster analysis. We want to find meaningful clusters of DRGs based on the dollar charges of our newly created cost category PCCR_OR_and_Anesth_Costs. To summarize, you should have 687 ish average costs in dollars and the analysis plan is to use them to create meaningful “clusters”. Among many methods of clustering that your software can perform, use the k-Means algorithm.

Obviously, you can set parameters in your software to identify any number of clusters. The optimal number of clusters could be determined by tracing the value of the Calinski-Harabasz f-statistics. Ask your software to report the Calinski-Harabasz f-statistics every time you run your cluster analysis. Try to cluster your cost data into 2,3,4, and 5 clusters and examine the f-stat each time to see where you have the best clustering of costs (i.e. where your f-stat is the highest). Take a note of your f-stats for this section however in the next section we only focus on solutions with 3 clusters.

Now consider your solution with only 3 clusters. Can you graph your costs and somehow color code your cost points to reflect the 3 clusters? You can use any tool for graphing your clusters. I used Excel and did it for 4 clusters just as an example and put it at the end of this section for you. You can get some ideas as how to do yours with 3 clusters. Or you can use any other methods as long as you can visually differentiate between your own 3 clusters. I am sure you will depict it prettier than mine 😊

Interpretation and making sense of your clusters:

This is where the things get exciting, sort of, or maybe not! First a bit of bad news ☹; more often than not machine-suggested clusters are not intuitive enough for users to grasp. Here are some typical reactions demonstrating the frustration of other users who have tried the AI:



“...I wonder why these X number of hospital DRGs are clustered together and the other Y hospital DRGs in another cluster and the other Z number in the third cluster? Why did my Python code decide to group them this way and why not in another way? They all look stupid and nonsense; I cannot see any similarity/commonality within cluster members, nor any clear dissimilarity moving from one cluster to the next. I do not like what I see and cannot wrap my head around these nonsense clusters. I hate this whole AI thing since it does not make any sense to me! I need to explain to my client who gave me the data and big money as what these clusters constitute but I do not have a reasonable story ☹ ”

Example quotes from confused AI community!

This is in particular the case when you want to cluster the data based on more than one variable. Here I simplified the world for you to cluster your DRGs using only and only one single cost variable. I could have asked for more than one variable to ruin your break!

So at this point your entire team needs to use its rich domain knowledge and come up with some good theories as what are the common properties of the hospital inpatient DRGs of each cluster that make them relatively similar or closely related to one another, and on the other hand, relatively dissimilar to

other clusters' DRGs. And beware not to make the very common mistake that most people do: do NOT focus on the costs and their magnitude within and between clusters more than what you have to, rather spend most of your brain on understanding the common properties of the DRGs. You are clustering and classifying inpatient DRG admissions and not costs. You only use certain cost categories to help you make meaningful classifications (clusters) of the hospital DRGs so stay focus on studying and exploring the DRGs based on many properties that are available to you through the Excel code lists and other resources.

So in your team class presentation spend most of your time and effort on a good interpretation and characterization of the DRG clusters since rest of the technical stuff are all easily done by your laptop! Human brain and thought process is vital to make sense of AI results, and that is a where the most focus of this exercise is.

Good luck!

Appendix:

Graph 1: An example of graphic presentation of the clusters (4 clusters here)

