HW3: Insurance Claims Data & Analytics

Due: TBD by TA

Context:

This homework lays out some analytics on the most important database in market for health, the Insurance Claim Data also known as Administrative Claim Data. We use an All Payer Claim Data avialbe to us from year 2016. The databases are rather large and will be made available to you by your TA. These databases consist of inpatient discharge data, outpatient procedures and services data, and emergency department data. Each data set includes:

- Case-specific diagnostic discharge data
- Some widely specified socio-demographic characteristics of the patient. Remember that details on demographics and addresses are not reported to protect patients' privacy as much as possible.
- 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. Note that these are what hospitals have charged and does not necessarily mean they would receive the exact amount from the insurance.

We are using the year 2016 databases and codebooks, also available to you thru this HW and along with data files.

Description of Data Files

- The Inpatient file reports the details of a hospital admission during the year based on the date of the discharge from the hospital being in that year. Each row represent one hospitalization and is assigned a Unique Identifier (variable Uniq in the file).
- Outpatient file reports the details of an outpatient encounter that took place in a hospital but not an admission to the hospital bed because the nature of the encounter was an outpatient service rather than a serious Inpatient hospitalization. Each row represents one encounter and is assigned a Unique Identifier (variable Uniq in the file). This Uniq identifier must be different and is in fact different from the Uniq numbers assigned in the Inpatient file. In other words none of the Uniq numbers used in the Inpatient file are used in the OP file vice versa.
- The third data file for each year includes only records relating to the Emergency Department (ED), and is extracted from both the Inpatient and Outpatient data files; it duplicates some records contained in the inpatient and outpatient files, because when a patient comes thru ED door he/she eventually triggers either an OP encounter or an IP admission in the hospital therefore its Uniq ID is maintained in the IP or OP file.

- The fourth data file is the Revenue Code file, which includes one record per revenue code for each discharge in the Inpatient, Outpatient, and Emergency Department files. Each discharge may have from one to many records in the revenue code file. The link between each of the IP, OP and ED files and the Revenue Code file is established using the Uniq variable.
- All data files are plain text format with values separated with comma and the name of the variables listed in the first row.

Question 1: Patient vignettes -- Files to use: Emergency Dept + Inpatient + Revenue Codes

We already had a patient vignette in class for the 75+ years old lady with infection problem who stayed in the hospital for more than 40 days and was discharged home afterwards. Her charge by the hospital was over $67K if you remember (Uniq ID = 254).

For this homework, please write a telling story for each of the following patients using their UNIQ identifier:

UNIQ: 507033, 40436, 859382, 1585831, 200760, 3692, 690326

Format: A comprehensive story is one that includes narratives on patient (pt) origin (admin source), admission type, the serving hospital, pt's demographics, insurance coverage, a nice picture of pt's diagnoses, discharge narrative, length of stay (number of days stayed) and of course the costs, and details of the services (nature and prices) according to Revenue Codes. A good story is similar to the work of detectives: as you scan thru the rather long list of up to twenty DX codes, you want to focus on the important facts that shed some light on pt's journey rather than just listing the DX codes one after the other with no meaningful connection between them.

200760: The patient, served by University of Vermont Medical Center, was from non-health care facility point with emergent health condition. She, a 18-24 years old young lady, had just gone through a motor-vehicle accident. She accepted diagnosis in the medical center where it was found that displaced fracture of medial malleolus of left tibia and unspecified fracture of shaft of left fibula. During the diagnosis process, doctor also diagnosed that the patient also had major depressive disorder and gastro-esophageal reflux disease. After that, doctor gave her a reposition on left tibia and left fibula with internal fixation device and she stayed in hospital for 4 days and was discharged home afterwards. Her charge by the hospital was over $49K, whose costs included Room & Board ($6768), Pharmacy ($1387.36), Medical/Surgical Supplies: Sterile supplies ($409.96), Medical/Surgical Supplies: Other implants ($10696.77), Laboratory-Clinical Diagnostic ($610.08), Radiology Diagnostic ($833.28), Operating Room Services ($14055.8), Anesthesia ($2590.07), Physical Therapy ($138.63), Physical Therapy: Evaluation/re-evaluation ($350.34), and Emergency Room ($2582.43). The costs were covered by her commercial insurance.

Hint: The information about most of these case study pts are found in the Inpatient file. However if the patient pathway begins from the Emergency Department then you want to start there first before you move onto the Inpatient database. As you get to the sections to talk about what is done for the patient,

given his/her diagnoses, then you want to link the ED and/or inpatient files with the Revenue Code file where all the details of the services rendered to the patient are listed. In explaining major services rendered to the patient use the REVCODE variable. The REVCODE list is found in the tab named "REVCODE" of the Excel file "*REVCODE_FILE_LAYOUT_and_CODES*". You need to link (Join if you use SQL) the Inpatient file to RevenueCode file using the variable Uniq. This is a synthetic unique ID assigned by computer to each hospital admission to facilitate following the same patient across different files.

Hint: Each hospital admission has one and only one Uniq ID in the IP or ED file but since each admission is associated with multiple services you will find the same Uniq ID repeated multiple times in the Revenue Code file. All the services in the Revenue Code file are listed under the same UNIQ ID. For example, the hospital admission identified by Uniq = 254 in the Inpatient file, has 301 services listed in the Revenue Code file showing all 301 service items rendered to the patient during that specific hospitalization.

Hint: If a patient comes to hospital multiple times during the year, each admission is identified by a new Uniq ID. Which is to say that the Uniq ID is used to identify unique hospital admission and not the patient.

Strongly Recommended! Make sure you all participate in this section of the assignment. For example by assigning at least one patient to each team member. In my experience you can only learn about the complex yet exciting world of the insurance claim data and analytics, if you undertake the task yourself with minimal guidance from others. I however do not grade the team submission by individual student's names and my suggestion is meant just as a reminder to get everybody engaged in the learning process.

Question 2: Service and Cost Profile of Major Insurances -- Files to use: 2016 Inpatient

Inpatient hospitalizations are identified and priced and paid by DRGs (Diagnostic Related Groups). So each hospital admission has a DRG. To learn more a PDF file is also included in this package explaining how the DRG classification system works. You can see the list of DRGs in the Excel file "*FILE_LAYOUT_and_CODES*". DRGs are classified in Major Diagnosis Groups or MDCs. You can find the name of the MDCs in the same Excel file. This question uses the MDC as a high-level classifier for an aggregated view of hospital admissions rather than hundreds of individual DRGs.

Identify the three major insurance payers: Medicare, Medicaid and then combine all hospital admissions of the two major commercial insurances of "BLUE CROSS" + "COMMERCIAL INSURANCE" and call this combined last category as **Commercial Payers** for this question. For the 3 insurances, create a cross tabulation with the MDC categories as row heading, and the name of the 3 insurances as the column headings. The cell values of your X-tab would be the sum of the dollar value of the charges for the selected MDC for each of the 3 insurances. Turn all your dollar values to $Million and round the value to drop any decimal points so you do not overwhelm your x-tab presentation with so many multiple digits figures. Drop all non-classified, unknown, and missing value rows from your x-tab and stay focus on the known MDCs only.

After presenting your x-tab in your report draw 3 pie-charts, one per insurance, to present the graphical view of the "inpatient services portfolio" for each insurance. Each pie chart starts with the largest MDC's share in percentages of all costs as the first slice at 12:00 o'clock.

Discuss the differences in the service portfolios across 3 pie-charts and try to relate the differences in the portfolios according to the demographics of the patients (age and sex is enough). You already know how the demographics of populations vary across the three major insurances yet you want to validate your assumptions using the real data. Your analytic writing can look like: from the pie chart of Medicaid we are observing that the top-5 MDCs are X, Y,Z,W, and Q and the reason MDC X is costing the Medicaid so much money is because Medicaid covers members with so and so demographics. However in Medicare the MDC so and so eats the largest portion of the money and that is because Medicare population is so and so .

Note: A good chunk of the grade for this question goes to your analytical story as backed by your x-tabulation figures and the pie-charts.

Question 3: Examining the enormity of the health crisis related to illicit drugs and prescription opioids use/abuse/overdose -- Files to use: Emergency Dept. file

Drug overdose has become a national health crisis in the US and globally. It causes many unnecessary costs and losses, financially, socially, economically, and most importantly in terms of losses of lives and/or quality of lives of the addicted individual, the family, friends, and the community. Many of us have been impacted by this crisis, one way or another, in our families, friend circles, or our communities. The governments and insurance companies are demanding the healthcare providers to pay more informational attention in submitting the drug abuse related claim data.

Not surprisingly, most of the drug abuse cases are brought to hospitals in a rather emergency manner so the proper database to study this question is the ED claims data. In order to study the drug abuse related cases the ICD-10 classification has assigned the entire blocks of T40 to T43 to such cases. Here are two ICD-10 code examples:

T401X1A Poisoning by heroin, accidental
T424X2A Poisoning by benzodiazepines, intentional self-harm, initial encounter.

Create a database that reports all the details form the ED file for every emergency department admission that has identified the pt with at least one drug abuse related ICD-10 code (i.e. any match with the entire code block including T40xxxx, T41xxxx, T42xxxx, and T43xxxx). Remember that these codes can appear in any location from DX1 to DX20. So do not assume that the doctor or nurse has recorded those drug use related ICD-10 codes only in the DX1 position. Also for some patients, multiple codes from the drug abuse block could be reported. This is because providers are required to report all details with regard to drug usage. However only one code from the block is enough for you to identify the patient as a drug use/abuse ED visit. If your data retrieval codes (SQL or any other search code) are

designed and implemented correctly, you will see slightly over 2,000 cases of ED admissions for drug use/abuse. If you are seeing fewer or more please debug your code and run again.

Once you have created the drug use/abuse analytical database answer these questions:

- How many ED visits exactly have been diagnosed as drug user/abuser?
- There is a myth that the drug use/abuse has been a male problem and that women have much better protection measures to stay away from drug use/abuse let alone overdoses severe enough that lead to an ED admission. Can you check if your data supports this gender bias myth?
- Tens of millions of dollars reportedly were spent on drug use related cases that year alone. Can you find the exact dollar amount for your identified patients in this question? Of the three insurances in Question 2, what was share of each of the total payments?
- Recent breakthroughs in the dark side of chemistry of drug development have done nothing but damage to humanity. The use of synthetic narcotics is rising alarmingly in part due to the marketing campaign for such meds. On the other front the public is ill informed of the danger of those new generation of lab created drugs that are supposedly improving brain's performance (read more here or here). Use the ICD-10 codes of T404xxx and T4362xx to identify only a small sample of such patients. How many of patients have been brought to ED for diagnosis related to synthetic narcotics or amphetamines?
- Name the 3 zip code regions with the highest numbers of drug use/abuse cases.
- What are the 10 most common diagnoses of drug use/abuse?

Hint: Remember that for this last part of this question you can have multiple drug use/abuse ICD-10 codes for some patients. See the patient with UNIQ = 19314 for example. He's been diagnosed with two drug use/abuse codes of T40605A and T4275XA, among a long list of other diagnoses. So if you pool all the reported codes you must get the number of all reported codes for all patients higher than the number of ED visits. Again, you need at least one code to find the patient to be a user/abuser, but then you need to pool all codes for all patients to answer this section of the question.