Assignment 3 – Analyzing Electronic Health Record data
Analysis of Large-Scale Data
2025

For this assignment, you will be using the de-identified EHR data from Beth Israel Deaconess Medical Center in Boston (MIMIC IV). **Download the files from the Homework 3 page on Canvas to begin the assignment.** As a reminder, this data encompasses all the hospitalizations and emergency room visits at BIDMC for 100 patients over 6 years.

The goal is for you to practice handing administrative data, as well as some of the Stata and analytic skills we have discussed in class. Remember that administrative data *is not collected for research*; we only use it for research as a *secondary purpose*. It can be messy and complex. Some other things that may be helpful to remember:

- Each subject can have more than one hospitalization record
- To de-identify the data, BIDMC converted years to be fake years that are far in the future (e.g., 2157). Don't let this confuse you – treat them like you would normal dates/years.
- regex(), substr(), and the date management functions you've learned may come in handy

One last note: while these data are real patient data, the small sample size and de-identification make it a little hard to come up with perfectly realistic exercises. After you turn in the assignment next week, we will debrief about what is and isn't realistic about this exercise.

1. Using the **admissions** file, create a variable that will capture the length of stay (LOS) for the hospitalization, measured in calendar days. You will need to convert the dates from strings to numeric/date variables. Create a histogram of LOS.

2. Now merge in three additional files: **patients**, **drg**, and **acuity**. Recall that DRGs are a summary of the hospitalization based on procedure and diagnosis information. Please recode any missing DRGs to be 9999. We will treat the unknown DRGs as their own category for the purposes of this assignment. Acuity is a value from 1 to 5 based on the triage assessment of the patient when they arrive, where 1 is most acute (urgent/dire) and 5 is least acute.

What percent of all hospitalizations are for Septicemia & Disseminated Infections (720)? What is the mean and standard deviation of the acuity score (to 2 decimal places)?

3. Now you will create and clean some of the demographic and covariate fields, so that they can be used in regressions. Calculate, create, and/or clean the following fields: age (continuous), race/ethnicity, insurance, and admission type. For race/ethnicity, convert it into 4 categories: Black, White, Hispanic, Other/Unknown. Please create a frequency table for your simplified race/ethnicity variable.

Hint: For creating race categories, regex may come in handy.

Hint: You will need to calculate age yourself.

4. Some patients are missing acuity values. Please create an imputed version of acuity, based on age, race/ethnicity, DRG, and admission type. What is the mean and standard deviation of the imputed version of the acuity score (to 2 decimal places)?

5. The first two characters in the admitting provider's ID are their group identifier (i.e., a group of providers who work together on a team, such as group "P0"). We want to compare the length of stay for each provider group. First, create a variable that captures the provider group's ID (e.g., "P0"). Using a regression, calculate the unadjusted means in length of stay by provider group and include a table below. Which provider group has the shortest unadjusted mean length of stay? Do any groups have a mean LOS that is significantly different than group P0?

Hint: substr() may come in handy.

6. Now calculate the risk-adjusted mean length of stay for each provider group. The characteristics you should adjust for are age, insurance, DRG, your imputed acuity, and a binary variable flagging whether acuity was missing. Include a table of the adjusted mean LOS by group below. Do any groups have an adjusted mean LOS that is significantly different than group P0? Finally, briefly describe how the risk-adjusted values differ from the unadjusted values and why they differ.

7. Now open the **diagnosis** file in Stata, which contains all the diagnoses listed on the EHR for each ED or hospitalization event. (You do not need to merge it to your existing file.) For each subject, calculate the total number of diagnoses that they've been given over the whole data period. What is the maximum number of total diagnoses observed per person?

Hint: don't count duplicate diagnosis codes.

8. We often want to create a version of administrative data sets that have <u>one row per patient</u> with each of their historical diagnoses in the columns. What Stata command would you use in order to get this diagnosis file into the desired format? Below, write some example code you would use to do this and annotate the code or explain your logic in words. (you don't need to actually run it or get it 100% right since these data are a little complex, but we want you to be on the right track by knowing the appropriate command(s) and variable(s)).

Hint: you may want to run the following line of code, which will create a variable numbering each of the diagnoses within each person.

```
bysort subject_id: generate dx_id = _n
```