Capstone Project Data Collection Overview

1. I explored three datasets relevant to healthcare utilization, which is the content focus of my capstone project. More specifically, I explored the detailed documentation available from the Healthcare Cost and Utilization Project (HCUP) concerning their National Inpatient Sample (NIS) database, Nationwide Emergency Department Sample (NEDS) database, and Nationwide Readmission Database (NRD). I chose to focus on HCUP data because it is the largest publicly accessible collection of hospital care data in the United States. I explored the documentation because this data costs approximately $200 per year per database (at student pricing), and requires an extensive application including training, a data use agreement, an indemnification agreement, and a contract concerning the responsibilities of the data purchaser; as such, I cannot link the datasets themselves but I have provided the documentation below. This documentation includes an overview and reports, restrictions on use, file specifications, a description of data elements (including summary statistics and frequencies), and more. I also needed to research how to work with ICD-10-CM/PCS codes concerning mental health; an example resource I examined on this topic is linked below.
    a. https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp
    b. https://www.hcup-us.ahrq.gov/db/nation/neds/nedsdbdocumentation.jsp
    c. https://www.hcup-us.ahrq.gov/db/nation/nrd/nrddbdocumentation.jsp
    d. https://www.who.int/classifications/icd/en/bluebook.pdf

2. No applicable code. I completed the extensive application and paid for the 2018 NRD. This dataset contains N = 17,686,511 records. “The Nationwide Readmissions Database (NRD) is a unique and powerful database designed to support various types of analyses of national readmission rates for all patients regardless of the expected payer for the hospital stay. The NRD includes discharges for patients with and without repeat hospital visits in a year and those who have died in the hospital. Repeat stays may or may not be related. The criteria to determine the relationship between hospital admissions is left to the analyst using the NRD. This database addresses a large gap in healthcare data - the lack of nationally representative information on hospital readmissions for all ages.” (HCUP). I chose to examine readmission because this is an important unifying topic for stakeholders and affects health outcomes, health costs, and demographic disparities. The large number of records and variables (K > 100) is ideal for machine learning.

3. I have saved the data to a password-protected 1 TB external SSD. The HCUP Data Use Agreement forbids me from sharing the data with unapproved parties and obligates that I maintain the data in a secured location, and there are substantial penalties (both civil and criminal) for violations of this agreement in the Indemnification Clause. For this reason, I have not uploaded the data to GitHub via the Git Large Storage Extension. While I cannot share access to the data, I will be able to share code and findings from this data including summary statistics, model evaluations, and so forth; therefore, the restrictions on data use will not impede my ability to collaborate with my mentor.