**PQHS 515: Health Care Analytics**

**Fall 2022**

**Homework 1**

**Due Date: September 23**

*Please submit via Canvas*

**Instructions**

The following code will help you read in the data:

libname in1 "/meta/db4/epbi515/homework/hmwk1/data";

**DATA** temp1;

SET in1.nis2014sample;

KEEP key\_nis hosp\_nis dx1-dx15 pr1-pr15 died female age totchg;

**RUN**;

This next line of code will show (on the .lst file) the first 10 records.

**PROC PRINT DATA =** temp1 (OBS = 10); **RUN**;

Addition sample code can be found in the SAS file “sample-procedures.sas” in /meta/db4/epbi515/homework/hmwk1

**Always check the log before reading the output! Important: DO NOT copy the sample file to your PC hard drive**

This homework introduces you to handling large administrative databases using SAS on unix. We shall use the Healthcare Cost and Utilization Project (HCUP) (<http://www.ahrq.gov/data/hcup/datahcup.htm>). HCUP is the largest all-payer collection of hospital inpatient care statistical information in the United States. The data you will use in this homework is a 10% sample of the 2014 HCUP data.

Use the variables key (patient ID), hospid (hospital ID), dx1-dx15 (15 fields of diagnostic codes), pr1-pr15 (15 fields of procedure codes), died (1/0 indicator variable of whether the patient died in hospital or not), female (1/0 indicator variable for sex), age (age at hospital admission) and totchg (total charge for the admission).

**Questions**

1. How many observations and how many variables does the file contain? (Hint: PROC CONTENTS) (2 pts)  
     
   707,177 observations and 36 variables
2. Find the number female admissions and the number of people who died in hospital in the dataset. (HINT: PROC FREQ) (2 pts)  
     
   number of female admissions = 405,304  
   number of people who died = 13,445
3. Find the mean, minimum, maximum and median total charge for all admissions in the dataset. (HINT: PROC MEANS) (2 pts)  
     
   mean = 41,699.33  
   median = 22,385.00  
   minimum = 100.00  
   maximum = 4,960,758.00
4. Find the mean, minimum, maximum and median age of all male patients. (HINT: PROC MEANS) (2 pts)  
     
   mean = 49.55  
   median = 57.00  
   minimum = 0.00  
   maximum = 90.00
5. How many pregnant women had a low cervical Cesarean section (ICD9 procedure code ‘**741’**) as a primary procedure? (HINT: IF/THEN, PROC FREQ) (2 pts)  
     
   22,338 women
6. How many people were diagnosed with influenza (ICD9 diagnosis code **487.0**), in any diagnosis field? (HINT: ARRAY, DO, IF/THEN, PROC FREQ). (4 pts)

1,080 people

1. Find out how many hospitals had 1,000 patients or more (HINT: PROC FREQ, OUT, DATA, IF). (2 pts)

27 hospitals

Short answer

1. Based on what you have learned about the HCUP NIS data from this exercise and in class, which of the following research questions do you feel could be adequately addressed by this dataset, and discuss why or why not:
   1. What is the mortality rate among all tonsillectomies in the United States (hint: tonsillectomies are often outpatient procedures, but not always).

No, because HCUP consists of inpatient data, so we would be missing lots of data on tonsillectomies given that the procedure is often outpatient

* 1. How many inpatient coronary artery bypass graft (CABG) surgeries are performed in the US per year, and what are the total charges from these surgeries.  
       
     Yes, we can do a similar query as to question 5 and then subset the data to those patients, thus being able to then find how many are performed and the total admission charges
  2. What is the global prevalence of tuberculosis?  
       
     No, this dataset only accounts for the United States
  3. What proportion of patients receiving surgery for temporal lobe epilepsy had a previous surgery to treat epilepsy in the past two years?  
       
     No, even though we can see if a patient was diagnosed with epilepsy, we don’t know if a surgery being done is for that diagnosis (unless there’s a procedure code that denotes it’s treating epilepsy)
  4. What is the distribution of payer type (public or private insurance) among patients receiving inpatient treatment for head trauma diagnosis?  
       
     Yes, given that this dataset the largest all-payer collection, I would assume that they have a variable that tells a patient’s payer type which could then be looked at on a subset of data where all the patients had a head trauma diagnosis