HCUP data

ZHE ZHENG

2023-01-13

File directory and its contents

The parent directory is named **hcup-sid**. Inside of the direcotry you will find:

- HCUPCode. It contains the sas codes to read in the asc files from HCUP data.
- HCUPData. It contains the original asc data files from the zip files you decompress from CD.
- ParquetData. This is the final data files you will use for analysis. Each state has two files. One file contains the core data of all years, including diagnoses, date of birth, date of admission etc. The other file named chgs contains the charges information (useful for cost-effectiveness studies).
- SASCode. These are the codes that Iris and I (Gigi) wrote to convert all the asc files to csv files.
- SASData. This contains the output csv files from the SAS program.
- Rcode. This contains the R codes to convert csv files to Parquet files. It also contains the R codes to run exploratory analyses using Parquet data.

Process to clean the HCUP data

(1) Read in CD. You need to ensure your computer have at least 200G space for storing the data.

(2) Uncompressed the zip files by inputting the passwords for each year and each state.

(3) Copy the asc files to directory HCUPData.

(4) Run the *sid time series.sas* in SASCode to run the SAS codes in HCUPCode. This will load all the asc data into SAS environment. And then run the *ConvertCSV.sas* to convert all data in SAS environment into csv files.

(5) Within Rcode, run R code *0_2.CreateParquet_states.Rmd* and modify it to convert all csv files to Parquet files.

(6) Within Rcode, modify R code *1_2.exploratory analyses.Rmd* to analyze parquet data in R environment.