Number of medical records: 140

Number of patients with at least one diagnosis identified: 73

Number of diagnosis': 449

Number of distinct diagnosis': 375

Average number of diagnoses per patient: 6

**Please include a brief description of the technology / method you used for this analysis. (E.g. data cleaning steps, and choice of programming/scripting methods.) Also submit any code used to process the notes.**

I used Java to clean and organize the data. The code goes through each line in each file until it comes to a line containing the phrase “discharge diagnosis” or “final diagnosis.” Then it parses the text following that string until it comes to a blank line. It separates the text into diagnoses by periods, commas, new lines, and dashes. It stores each diagnosis and a count of the number of times that diagnosis has been identified in a map. Also, there is another map of common acronyms and the full text of those acronyms used to translate acronyms. I also use a set of strings to stop false diagnosis’, like the strings “primary” and “patient” from being treated as a diagnosis and stored in the map. I used the final map to write a table in a .csv file named Diagnosis.csv. The code also prints out the statistics to the console.