Instructions

We ask you to solve three short assignments involving analytical reasoning and programming. There is not necessarily a right answer to the questions being asked and a simple solution is preferred over a very complex one. We are interested both in your code and how you reason when solving an assignment like this, so please send us your commented code, including brief motivations of the chosen solution, in addition to your answers to the specific questions. You can format this in any way you chose and use whatever statistical tool you prefer, but please don't overwork it!

Data

You've received a folder containing data from a mock register linkage on a cohort of patients with rheumatoid arthritis (RA), a chronic inflammatory joint disease, who have started one of two treatments: *A* and *B*. The unique key between the datasets is *pnr*. The folder contains the following data files in CSV format:

1. **The Swedish Rheumatology Quality of Care Register (*clinical.csv*)**
   Data on drug treatment, and disease activity at treatment start. The disease activity measures are C-reactive protein (*CRP*), patient-reported pain (*pain*) and the number of tender and swollen joints: Tender Joint Count (*TJC*) and Swollen Joint Count (*SJC*) respectively. Treatment data includes: starting drug A or B (*drug*), the date of treatment start (*start*), the number of months of treatment follow-up (*fu_trt*), and an indicator of whether the patient stopped the drug or was censored at the end of follow-up (*stopped_trt*).
2. **The National Patient Register, outpatient part (*outpatient.csv*):**
   Dates and diagnoses from the outpatient component of the National Patient Register. *Hdia* is the main diagnosis, and *Dia01* to *Dia06* are the first 6 contributory diagnoses. Date of visit is *indatum*.
3. **The National Patient Register, inpatient part (*inpatient.csv*):**
   Dates and diagnoses from the inpatient component of the National Patient Register. *Hdia* is the main diagnosis, and *Dia01* to *Dia06* are the first 6 contributory diagnoses. Admission date is *indatum*, discharge date is *utdatum*.

Assignment 1

Briefly describe the data (numbers of observations, variable types, et cetera). **Make a (brief!) note of things you notice in this description that may be relevant for using the data in a research project.**

Assignment 2

If we wanted to run a study comparing the outcome of drug A versus drug B, we would worry about confounding by factors associated with the choice of drug. One such factor could be the medical history of the patient. In this assignment, your task is to assess if there is a difference in the burden of disease in the years leading up to treatment start.

For each subject starting drug A or B, create two new variables by counting the number of times *in the five years before treatment start* that they:

1. Had an inpatient visit
2. Had an in- or outpatient visit where either the main or contributory diagnoses started with any of: I200, I21, I22, I23, I24. (A visit with multiple diagnoses should only be counted once)

When deciding the date of an inpatient visit, use "indatum". Count each out- and inpatient visit, even if recorded on the same date.

Using any test or similar you deem necessary, assess if the two counts differ by drug A vs B

Assignment 3

Using any test or similar you deem necessary, assess if any of the variables *drug*, *CRP*, *pain*, *TJC*, or *SJC* predict stopping treatment