Final Project Data Memo

Vasyl Ostapenko (774 970 8)

April 06, 2022

Dataset Overview

For my dataset, I will be using the Vaccine Adverse Effect Reporting System (VAERS) dataset for calendar year 2021. The VAERS database is a joint effort by the FDA and the CDC to provide a system for reporting both minor and serious events related to vaccines. One could find further information as well as the data at the following link: https://vaers.hhs.gov/data.html. The data is split into three tables, called "data", "symptoms", and "vaccine", respectively. Each row of every table is a separate event, with a unique ID attached. The unique event ID could be used to map between the three tables. The "data" table also provides information on the patient, their symptoms, and their treatment. The "symptoms" table additionally lists symptoms related to each event in further detail and codes them in the internationally-accepted medDRA format. Finally, the "vaccine" table gives further information on the vaccine related to each event. There are about 750,000 - 1,000,000 observations in each of the three tables. However, I anticipate the amount of useful observations to drop due to my more focused research question. In terms of predictors, I anticipate to use anywhere from 15 - 30 columns of the dataset. I will be working with factors, continuous variables, dates and times, as well as some text data. Each event is reported by a medical provider and every patient's situation is unique and thus there is missing data. I plan to do a combination of: dropping observations with missing values; removing columns with mostly missing values; imputing missing data.

Research Question Overview

For my question, I would like to study the relationship between the features of each COVID-19 vaccine adverse event and the symptom of myocarditis (inflammation of the heart muscle). Another sub-question I would like to consider is what is the difference between the three main COVID-19 vaccines (Pfizer, Moderna, JJ) in their relationship to myocarditis. Concretely, I will pull as many features related to the patient, their vaccine, and their other symptoms as I can. I will then build a predictive model to classify whether each event will include the symptom of myocarditis. I believe that other symptoms for each event will be very useful in predicting myocarditis. Thus I will use clustering analysis to get an idea of what symptoms fall into the same group as myocarditis.

Proposed Project Timeline

I already have the dataset loaded. I will begin EDA, data cleaning, and feature engineering sometime next week. My plan is to spend a few hours each week working on the project until the deadline. I definitely believe that cleaning the data and extracting features will take the longest amount of time. On the other hand, building a few different models and training and testing them should be a piece of cake.

Potential Concerns

Extracting features from the data will be the most difficult part. The data is not standardized in some parts due to the decentralized reporting. The data also includes text information on symptoms, allergies, etc. which will have to be processed.