For this challenge, we were given 23 datasets from those datasets we were able to create 4 combined datasets Locations, Observations, Conditions and Person. Locations consists of the provider, location and care site which were joined by data partner id and person id. Observations consist of observation period, observation and visit occurrence. To join the observations dataset, we used the data partner id and person id and to combine the visit occurrence dataset we included the visit occurrence id. Conditions consist of condition occurrence and condition era which were combined using their respective person id and data partner id. For the person dataset we used the person and Long COVID Silver Standard file and were joined using the person id and data partner id.

In the person dataset we had the 92% missing values which consisted of provider id and care site id, after combining the Person and Locations dataset we were able to find the missing information from their original datasets and fill those values. The time to pasc column consists of 80% missing values which correspond to a patient not having long covid symptoms prior or after their covid index date, to fill these missing values we will assume 0 days if they did not test positive for long covid symptoms after their covid index. The locations dataset has missing values for the state and zip code of the locations, we will not be using these columns in our model dataset. In the Condition and Observations dataset we have 19% missing values in the visit occurrence id which corresponds to the unique visit identifier as well as the visit concept id which has 19% missing values. The visit occurrence id column was used to merge datasets but will not be used in the model dataset. The visit concept id will also not be used since we are only using the observation and condition concept id in our model since they differ only due to a provider classification of the condition. We can assume that what condition a patient has can differ due to the interpretation of symptoms by the provider. We can also assume that although a patient can have symptoms found in long covid it does not necessarily mean that their condition relates to covid at all. The CDC defines long covid to have a large range of new, returning or ongoing health problems that with testing positive for covid can result in a patient developing long covid. Considering that at least four weeks after infection of covid is the start of long covid conditions can be identified we will be using pasc code after four weeks to determine long covid patients. Currently in our dataset we have a total of 57, 672 patients of which we found 9,031 patients who have the PASC code after four weeks of infection. These will be used to predict the outcome of a patient developing long covid.

Missing values:

Person Dataset:

"time\_to\_pasc": 80%

"care\_site\_id": 92%

"provider\_id": 93%

Locations:

Provider\_id: 17%

Observations:

Visit\_occurence\_id: 19%

Visit\_concept\_id: 19%