Progress Report 2

### What progress has been made so far, and how it lines up with the plan and timeline submitted.

We have finished creating an account on allofus. Mica was not able to verify identity during the account creation process in the first week. It took another half a week to get the verification figured out for her account. Both account has proof of identity completed, finished the allofus training, and passed all the tests necessary to access the data on allofus.

A cohort was created that included patients that have taken the survey or have conditions. We selected the data to include all available columns that we could use in each of the following tables (conditions, surveys, and patient demographics). Then, we created a workspace where we can write python code. We changed which conditions we are looking into to be acute myocardial infarction (heart attack), rheumatoid arthritis, diabetes, and sleep apnea. We selected patients that had a icd 10 or icd 9 code for one of the conditions listed previously.

We were able to merge the tables together so that it is one patient per row and to have the questions assigned to specific columns instead of randomly merging together. We dropped columns that we did not think were relevant such as code ids, some source name, source values, ect. The data has been preprocessed into one-hot encoded matrices, with train/test and data/target splits.

Additionally we have created two rough models, one basic fully connected feed-foward neural network and one Bayesian neural network. Currently the full connected model only has two layers, and the Bayesian only has one. Both have initially only been trained to predict one disease’s occurrence. The intent is to make sure that the models are working how we want them to and mostly tuned before scaling up. This is to cut down on development time.

Currently we are a little behind schedule. At this point we wanted to have the fully connected network as well as the Bayesian network finished and generating some initial results, however we have not. This is largely due to us both being distracted with other work, and not putting a higher priority on this project resulting in less work being done at regular intervals.

### What's been done since the last progress report.

Since the last report we’ve completed the data preprocessing and the initial construction of the models. To reiterate the relevant section from the previous part of the report:

The data has been preprocessed into one-hot encoded matrices, with train/test and data/target splits. Additionally we have created two rough models, one basic fully connected feed-foward neural network and one Bayesian neural network. Currently the full connected model only has two layers, and the Bayesian only has one. Both have initially only been trained to predict one disease’s occurrence. The intent is to make sure that the models are working how we want them to and mostly tuned before scaling up. This is to cut down on development time.

### What challenges or obstacles have arisen.

The Bayesian network has been harder to finish than initially anticipated. The library that we are working with is not entirely user-friendly and this has led to some slowdowns. Additionally both of us are highly distracted by our qualifying exams and research projects, and have not dedicated time at regular intervals to this project.

### What the next steps are – what do you expect to complete between now and the next report?

The next steps are to finish both models, tune their hyperparameters, and scale up the models to predict all of our diseases. We want to have run multiple experiments and generated some solid results for our report. Additionally, the next report is the final presentation, so we need to have finished our background research by that point. We would like to have finished the final paper as well.