Our team received data from the Geisinger office in Frankfurt. The data was saved in a thumb drive and mailed over to our team, which took about a week. Our data was saved locally to our teammate’s laptop, and from there we initiated the data cleaning process. The first problem we encountered was the size of the dataset. This clinical dataset contains 7 large CSV files spanning the years of 1995-2017, and is composed of a lot of clinical vocabularies that our team were having trouble understanding with. However, after meeting and investigating into the “data dictionary” file, which was a summary of the whole database, we chose the direction of analyzing the correlation of depression and insomnia as well as other demographic factors including age, gender, ethnicity and more.

We had a meeting with Professor Ohayon and Professor Cristina from Stanford University in the middle of the semester. By that time we just finished cleaning, blending the data and building our first logistic regression model. After looking into our results in details, Professor Ohayon and Professor Cristina suggested us to also include the factor of anxiety to our data and further investigate the causal relationships among three: anxiety, insomnia and depression.

Later, our team focused on finding the correlation in depression, insomnia and anxiety as well as demographic factors using logistic regression and random forest classification. We verified that the relationships among depression, insomnia and anxiety are multi-directional. Furthermore, we confirmed research results onto the factors predisposed towards depression -- females, anxiety patients, or insomnia patients, for instance. The outcome of our project gives a general sense of which population are more likely to be diagnosed as depressed and will eventually help doctors identify a patient’s likelihood of getting depression and thus mitigate the effects of mood disorders.

Given that the current accuracy of both our models is around 70%, we are trying to improve the accuracy in the future such as adding patients’ social history (drug use, alcohol consumption, and smoking habit) to our model, and tuning the classifier parameters.