Milestone #5: Proposal of Future Work

Prompt: Provide a roughly 1-page outline of what approaches, models, etc. you would like to use for the final project results.

A take-away message from our milestone 4 is that although the results are not very promising, there is a lot of room for improvement. Also, we have noticed that there is a relationship between previous seizure incidences and possible future epilepsy recurrences; age may not play an important role in predicting epilepsy; the treatment, progabide, may reduce the recurrence of epilepsy; and linear regression model could possibly be a good start.

Our goals for the following one-week are:

1. **Further clean up and massage NCDS dataset:** while investigating the messy dataset, we noticed that due to the **inherent** property of longitudinal study, the span of the NCDS dataset is very long, thus, the measure and even the identification of epilepsy are not constant during the onset study period. Therefore, instead of focusing on epilepsy, we decided to include convulsion as well, moreover, we decided to use **"dataset 5565: NCDS: Childhood Data, Sweeps 0-3, 1958-1974"** to investigate the effects of early childhood epilepsy syndromes on adolescent recurrences of certain diseases. In this dataset, 1026 children had shown syndromes of epilepsy and convulsion before age 11, among them, 250 developed epilepsy or convulsion between ages 11 to 16. There are many risk factors that could be associated to the recurrence of epilepsy. Thus, our next steps are removing invalid entries, checking for possible correlations among variables, and visualizing the effects of key factors that we identified through various research papers.

2. **Models to run:** Based on our findings from milestone 4, we decided to run **linear regression model** to evaluate the variables associated with epilepsy and convulsion; **logistic regression model** to predict the recurrence of subsequent seizures after the first occurrence; and due to the fact that we have a high dimensional dataset, we will run **linear discriminant analysis** to reduce the dimension.

3. **Possible challenges to overcome:** There are 1765 variables; therefore, we need to deal with high dimensional data. We decided to manipulate our subject matter knowledge gained through reading research papers, and also make use of LDA, PCA, and lasso. Moreover, there are a lot of categorical and missing data that we have to consider, besides, possible highly correlated variables are presented, we will do our best to confront these challenges.