Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations

The paper aims to utilize big data analytics for Parkinson's disease, where it is collected and managed by Parkinson's Progression Markers Initiative (PPMI). Previous studies have examined the relationship of the disease and risk to trauma, genetics, environment, or lifestyle, and Prof. Draganski's team have integrated complex PPMI imaging, genetics, and clinical/demographic data. The paper emphasizes the collection of data of multiple sources to encompass the defining characteristics of Big Data, which include large size, incongruency, incompleteness, complexity, multiplicity of scales, and heterogeneity of data. Data characterization in this paper is uniquely spread into 3 main concepts: introduction of new methods for rebalancing data, use of wide classification methods for consistent phenotype predictions, and generate reproducible machine-learning classification that can incorporate new data as well as application for other neurodegenerative disorders.

Upon evaluations of several predictive models, some failed to generate accurate and reliable diagnostic predictions because of the nature of the databases and the strengths of the models did not match the nuances of the data used in this study.

We found this paper incredibly interesting because of the challenge of creating effective predictive models based on heterogeneous data. Building predictive models on heterogeneous data is incredibly difficult because a heterogeneous population or sample is one where every member has a different value for the characteristic you're interested in. A heterogeneous dataset would make it more difficult to reach conclusions as patterns that would be used to predict Parkinson's disease would be difficult to find. The scientists involved in the study indicated that they began their analysis by fitting their data to statistical models and later on moved onto vector-based machine learning approach, and it was interesting to see their process to better understand the data through analysis then better more through trialed statistical modeling, before arriving at using a machine-learning method which from the start would have been the obvious method of analysis to use for the type of predictors used in the study.

Because the Mimic database contains highly specific data to the ICU, a lot of the data necessary to pursue the analytics described in the paper is not available. In the paper there was exploratory analysis done on gender and their implications on past studies on parkinson's. The analysis from the paper proved that gender had negligible implications on getting Parkinson's. Using the mimic database, we queried the distribution of gender amongst admitted parkinson's patients, and we analyzed the

death rate of parkinson's patients in male and female patients. Although a more interesting query could be made with the age of the patients, because of the compliance to HIPPA, the age for many of the patients is altered for the sake of patient privacy. Due to this, analysis based on age would prove to be very difficult, especially since the population of patients with Parkinson's would be around the 89 year age which is in the age range of many parkinson's patients given that Parkinson's is a disease that affects 1 in every 100 people over 60 years of age and parkinson's patients have a lifespan of around 10-20 years after diagnosis. We figured that the variable 'expire_flag' would serve as a makeshift variable that could be indicative of age, operating under the assumption that most parkinson's patients die when they are more senior without factoring in the possibilities of early onset Parkinson's, and other reasons for death.

To acquire the data necessary to perform the paper's analysis we would have to search outside the ICU data provided by the mimic database. To build a predictive model based on PPMI imaging, genetics, and clinical/demographic data we would need to acquire this data from consenting patients from their general physicals, genetic studies, and from data collection methods that. But for the purpose of simplicity for the open query, the voluntary provision of basic physical information such as height, weight, bmi, and history of neurological diseases would be helpful to perform part of the analysis described in the paper.