Determining the number of OFF Periods per week for Parkinson's Disease patients

Adil Baran Narin, Abhishek Shirsat, Rahul Kamal Pandey

November 2019

Abstract

Parkinson's Disease is one of the diseases which remains unexplored and has a lot of unanswered questions associated with it. With the actual cause still unknown, PD cases arise spontaneously or are hereditary. So far, it has been established that the brain cells responsible for 'dopamine' (a hormone that regulates our body movements) generation die off in PD. There are many aspects to PD which still remain untouched. In recent years, Data Analytics has opened a new field of research in understanding the causes and symptoms of PD. In this paper, we discuss how statistical learning methods can be used to develop various predictive models. These predictive models are used to study the symptoms and their level of impact on daily activities and relate it to the number of OFF periods a patient has in one week. The model accuracy is assessed and the best fit model is derived for the dataset in consideration.

1 Introduction

National Institute of Neurological Disorder and Strokes states that Parkinson's disease (PD) belongs to a group of conditions called motor system disorders, which are the result of the loss of dopamine-producing brain cells. James Parkinson has been credited for almost 200 years with establishing the major clinical features of Parkinson's disease (PD), and we remain in awe of his observational and descriptive skills in characterizing so many of the core elements of the disorder that bears his name [15]. Section 1.1 outlines the causes, symptoms and available treatments for PD. In Section 1.2 we will discuss how data analytics has opened a new realm of research in answering the previously unanswered questions pertaining to PD. Machine learning is making huge strides in understanding various aspects of PD. Thus, applications of statistical learning methods in Parkinson's disease is a good read. In Section 2 we will discuss the dataset source, the initial exploratory data analysis, required data transformation and various regression models that can be used to make requisite predictions. We conclude by summarizing our findings from this project and discuss it's future scope.

1.1 Parkinson's Disease: An Overview

Parkinson's disease (PD) is a long-term degenerative disorder of the central nervous system that mainly affects the motor system. More than a million people in United states (10 million worldwide) are living with the symptoms of PD which get worse over time. At present, there is no cure for PD, but a variety of medications provide dramatic relief from the symptoms. The early stage symptoms include involuntary tremors, stiffness, acute constipation, humped standing posture and sleep dysfunctions. The advanced stage symptoms are dysautonomia (a dysfunction of autonomic nervous system affecting heart, digestive system and other organs), impaired balance of body, Gait (Dopamine deficiency), Dementia (memory loss) and reduced blinking rate. The diagnosis focuses on improving the health of Central nervous system which includes tests like cranial nerve examination, Deep tendon reflex tests and mental state examination. Since PD occurrence and development varies from person to person, the treatment is based on the patient's condition. The focus is on increasing the dopamine levels in brain and since it is a chemical, it cannot be given directly since it cannot enter brains. The most effective medication so far has been Carbidopa-Levodopa (CL), a chemical that enter the brain and gets converted to Dopamine. Dopamine agonists are

another prescribed medication that mimic the effects of Dopamine in brain. It lasts longer than CL and is used to smoothen the OFF-ON effects of CL. Mao-B inhibitors help to prevent the dopamine breakdown in brain by inhibiting the brain enzyme named Monoamine Oxidase B. In addition to these medications, Deep brain stimulation and Neuroablative Lesion surgery has been proved successful in recent years. Stem cell therapy is another potential cure but the research is ongoing. PD begins 5 – 10 years before any clinical symptoms appear. By this time, approximately half of the dopaminergic neurons are already affected. Non-typical signs like depression, pain, fatigue, etc makes the diagnosis even difficult. Currently, the diagnosis made by doctors to assess the severity level of PD is conducted using various methods based on several research domains, including cognitive deficits [3] [8] [14] [6] [4] [11], speech disorders [11,12,13], human stability [14,15], gait cycle and others. It is clear that the lack of specific test for Parkinson's disease makes it challenging to diagnose PD subjects.

1.2 Data Analytics and Parkinson's Disease - Literature Review

Big Data is making major strides in grossly expanding what we know about the disease—and thus making a cure more likely in years to come [12]. Nowadays, Researchers are getting help from the big data tools to enlighten chaotic situation of not just PD but also other topics of Healthcare. Contemporarily genomics and postgenomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms [3] That's why the data sets have heterogeneous, complex, large datasets that collected from the patients. Also, they are hard to manage and analyse with the traditional methods.[8] [8]. There are six characteristic of Big Data which denoted 6Vs. Those are value, volume, velocity, variety, veracity and verability. On the other hand, some outhers describe more than 6 features of Big Data[15]. However, Implementation of big data analytics to manage this types of data sets helps us to control the quality of the data, analysis of the data, modelling, and interpreting [14]. [5] developed a comprehensive protocol for end-to-end data characterization, manipulation, processing, cleaning, analysis and validation. They processed complex PPMI imaging, genetics, clinical, and demographic data. They used both model based and model free approaches for predictive analysis . Model based approach includes GLM, MMRM, GEE and change based models whereas model free approaches include forecasting, classification and data mining, KNN, Decision tree, SVM, K-Means. For validating they used n-cross validation. They observed that for better predictive analysis for an incomplete and heterogenous data, statistical rebalancing of cohort sizes yield better discrimination of group differences. Model based approaches failed to generate accurate and reliable diagnostic predictions. Many Machine learning based classification methods outperform model based approaches and they predicted Parkinson's disease really well in PPMI subjects. Sensitivity, consistent accuracy and specificity more than 96 **percent** was confirmed using n-fold cross validation.
In [10] the authors proposed classifying patients with PD and healthy control subjects using gait analysis through deterministic learning theory. This classification approach consists of two phases: a training phase and a classification phase. In the classification phase, a bank of dynamic estimators was constructed from all the training data. The results show that this approach achieves an accuracy rate of 96.39 **percent**. The paper's ultimate goal is to developing classification and predictive model for PD with utilizing PPMI neuroimaging, genetics, clinical and demographic data. However, the fundamental data science theory does not allow the Author use entire data raw for creating model accurately. Therefore, Author uses different data sets individually and combine them with each other. The authors have employed the statistical computing environment R for their model fitting, parameter estimation and machine learning classification.
[10] discuss different supervised machine learning models namely K-nearest neighbour (K-NN), Decision tree (DT) or Random Forest (RF) can be used. K-means, an unsupervised machine learning technique can also prove useful. Classification with K-NN involves two main steps: (1) a distance calculation (usually, Euclidean distance) is made between the new sample and all training samples; (2) the new sample is assigned to the majority class of the nearest samples using the K nearest neighbour selection. Clustering methodology can be used when we have a large number of predictors. Comparison of clustering algorithms on a merged dataset helps to determine the best approach. The types of approaches are k-means, k-medoids and DBSCAN. Clustering merged data into different number of clusters and evaluating the quality of produced clusters with the internal cluster validity metrics; SA (Rousseeuw 1987), DB (Davies and Bouldin 1979), and CH (Caliński and Harabasz 1974) lead to best suitable approach. Better clusters are marked with higher values of SA, CH and lower values of DB (Valmarska et.al 2018, para 2)(18) In order to select the relevant features,

SPEC algorithm (Zhao and Liu 2007) can be implemented. A plot of feature importance vs feature rank is obtained and the features to left of reference line are selected as most relevant. A cluster validation is implemented considering these selected variables. The results revealed that the merged data set (consisting of sums of attributes) produces better quality clusters than the data set reduced with feature subset selection [16] Parkinson's disease patients experience a whole range of symptoms, both motor and non-motor, and it is tougher for traditional clustering algorithms to separate them into groups of similar patients. The introduction of sums makes it possible to have a view of the overall status of the patients concerning particular sets of symptoms (i.e. motor symptoms, non-motor symptoms, autonomic symptoms etc.).
[7] discussed how deep learning algorithms can be successfully applied in medical field. Majority of these studies are based in use of convolutional neural networks for brain segmentation, tumour detection and diabetic rectinopathy. Long term memory algorithms are also used for general diagnosis of diseases. Deep learning algorithms can be used to investigate challenges in PD. The severity of PD is a measured using Unified Parkinson Disease Rating scale (UPDRS) which record the patient's various movements at regular intervals. Better understanding of UPDRS can help data analysts to formulate problems that use the capabilities of deep learning and reduce the manual effort spent to analyze the data. The data should include the labels of diagnostic elements for us to better answer different research questions. LSTM models require time-series type data. LSTM's can be trained to predict how the disease progresses for different types of medications using the UPDRS data that contains the record of disease for a PD patient over the years. Very little research has been done on deep learning algorithms applied to PD data. Diagnosis of this disease involves multiple features which include motor functions, dopamine concentration and many other clinical tests. Using just one feature is not enough since there are multiple brain related diseases with similar symptoms. The process of using deep learning algorithms to PD can be simplified using trained deep neural networks. The error rate for diagnosis of PD can be as high as 24 **percent** but most of the studies that yield this result include atypical parkinsonian disorder (APDs) like supranuclear palsy (PSP), multiple system atrophy (MSA) etc.

2 Data Description and Transformation

In this section, we discuss the dataset used in this study. The source of our dataset, the description of variables used to make prediction, the response variable, data transformation and basic data visualization is discussed in brief.

2.1 Data Collection

The data set used in this study is obtained from Michael J. Fox foundation website. Intel has teamed with the Michael J Fox Foundation to gather data from Parkinson's patients, with the aim of using that data to better understand the disease.(1). Through this paper, we intend to study the patients of different age groups. The factors in consideration are symptoms displayed by patients, the impact it has on their daily activities, the proportion of predictable and unpredictable OFF periods, the average approximate OFF days they experienced in a week and try to categorize them into types of severity. An OFF period is defined as the period when a the PD symptoms return for a patient on CL medication(2). CL is considered to the gold standard of prescribed medication for PD(2).

2.2 Variables in study, Data Pre-processing and Transformation

The variables into consideration include the age of patients, the duration of OFF periods, the method used to keep track of OFF periods, a variety of symptoms, the impact they have on the daily lives of patients and communication between patient and doctor for each patient. We are going to focus on the age, tracking method used by patients, the symptoms and impact it has. The first step is thus going to be eliminating the variables we do not need. Since these omitted variables are not correlated to the other features, it is not going to affect our analysis. The features mentioned in the data set have a value response that carries a different description. They are as follows:

Table 1: **Variables and Values description - Part 1**

Variable	**Value**	**Value description**	**Variable description**
OFF1ExpStart	1	Less than 1 year	How many years ago did you begin to experience OFF periods?
OFF1ExpStart	2	1 to 5 years	
OFF1ExpStart	3	6 to 10 years	
OFF1ExpStart	5	I don't know	
OFF1ExpStart	4	Greater than 10 years	
OFF1WeekNum	2	1 period per day	Over the last week, on average, how many OFF periods do you Experience in a typical waking day?
OFF1WeekNum	4	3 periods per day	
OFF1WeekNum	3	2 periods per day	
OFF1WeekNum	1	No periods, zero	
OFF1WeekNum	5	Greater than 4 periods per day	
OFF1WeekNum	6	I don't know	
OFF1WeekDur	6	I don't know	Over the last week, on average, what is the typical duration of each OFF period?
OFF1WeekDur	3	Between 30 and 45 minutes	
OFF1WeekDur	2	Between 15 and 30 minutes	
OFF1WeekDur	1	Less than 15 minutes	
OFF1WeekDur	4	Between 45 minutes and 1 hour	
OFF1WeekDur	5	Greater than 2 hours	
OFF1UnpredProp	1	0%	What proportion of your OFF periods come at unpredictable (i.e. unexpected) times?
OFF1UnpredProp	3	25-50%	
OFF1UnpredProp	2	Less than 25%	
OFF1UnpredProp	5	I don't know	
OFF1UnpredProp	4	Greater than 50%	
OFF1Predict	3	Neutral	If the timing of your OFF periods were more predictable, how much would that lessen their impact on your life?
OFF1Predict	1	Very much	
OFF1Predict	2	Somewhat	
OFF1Predict	4	Not at all	
OFF1Track	2	No	Tracking method used by patients
OFF1Track	1	Yes	
OFF1TrackPaper	1	Checked	
OFF1TrackElect	1	Checked	
OFF1TrackOth	1	Checked	

The information about symptoms is such that it first has a value - 1,2 or 3 assigned to it which means Yes, No or Unsure respectively. The description for each of the symptoms is as below:

Table 2: **Variables and Values description - Part 2**

OFF1SympFatg	Fatigue
OFF1SympSleep	Sleepiness
OFF1SympTrem	Tremor
OFF1SympStiff	Stiffness
OFF1SympSlow	Slowness of movement
OFF1SympGait	Change in gait/walking
OFF1SympFall	Increased falls
OFF1SympHand	Difficulty with hand coordination
OFF1SympSwall	Difficulty swallowing
OFF1SympSpeak	Difficulty speaking
OFF1SympBreath	Trouble breathing
OFF1SympNaus	Nausea
OFF1SympPain	Pain
OFF1SympAnx	Anxiety
OFF1SympIrrit	Irritability
OFF1SympAgit	Agitation or restlessness
OFF1SympMotiv	Loss of motivation
OFF1SympSad	Sadness/depression
OFF1SympSocial	Social withdrawal
OFF1SympFlash	Hot flashes
OFF1SympSweat	Sweating
OFF1SympAppet	Loss of appetite
OFF1SympBladd	Change in bladder function (e.g. urgency, incontinence)
OFF1SympThink	Difficulty thinking
OFF1SympOth	Are there other symptoms not previously listed that you experience during OFF periods?

For each of the symptoms mentioned in dataset as above, we have a rating from 1 to 5 where 1 means no impact and 5 means severe impact. If for a particular patient, the information on symptoms is 'Yes', we have a rated value associated to it. If the value associated to a symptom is 'No', the rating is left Empty. Thus, we have plenty of NA values in our dataset which makes the analysis tedious. To begin with, We determine a set percentage for each of our variable columns having NA values. For the entire range of patients, we delete the columns with more than 90 **percent** NA values and similarly for rows, we delete those with more than 70 **percent** NA values. Secondly, if the patient has not experienced a particular symptom, we assign a rating of 1 to it which means no impact. For the variable 'Other symptoms, we assign a value of 3 which for just this variable means 'Not sure'.

The impact variables from dataset are as below:

Table 3: **Variables and Values description - Part 3**

OFF1ImpactRateCare	Please rate the impact of X experienced during OFF periods by the person you care for with Parkinson's disease.
OFF1ImpactDay	In general, how much impact do the OFF periods have on your daily life?
OFF1ImpactPhys	Physical activity
OFF1ImpactLeis	Leisure/hobbies
OFF1ImpactEmpl	Employment
OFF1ImpactRelat	Relationship to care partner
OFF1ImpactFriend	Friendship
OFF1ImpactHouse	Household tasks
OFF1ImpactDrive	Driving
OFF1ImpactGroom	Self-care/grooming
OFF1ImpactIndep	Independence
OFF1ImpactComm	Communication
OFF1ImpactLeave	Your freedom to leave the home
OFF1ImpactActiv	Scheduled activities
OFF1ImpactFrust	Off periods frustrate me
OFF1ImpactAnx	OFF periods make me anxious
OFF1ImpactScare	Having OFF periods is scary
OFF1ImpactEsteem	Having OFF periods has hurt my self-esteem
OFF1ImpactEmbar	OFF periods make me feel embarrassed

Similar to symptoms, the impact is rated on a scale of 1 to 5. Similar to symptoms, For impact variables, we assign a value 0 to NA terms which means 'We do not know'. We have thus transformed our dataset from 127 variables to 66 variables for 572 patients. Although using multiple features to achieve accurate classification is common, increasing the number of features might lead to dimensionality which results in a degraded classifier performance and increases the computation time and model complexity. The feature selection step involves finding the most relevant subset of features from the original dataset and eliminating the inappropriate ones. In this study, we eliminate the features pertaining to the communication between the patients and doctors/medical advisors.

3 Data Visualization

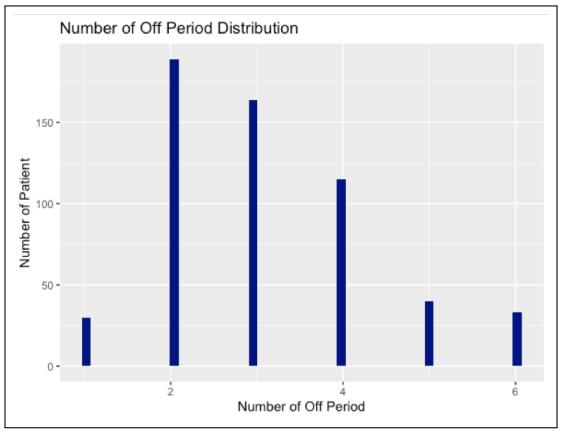

Figure 1: Total number of patients based on number of OFF periods experienced per week

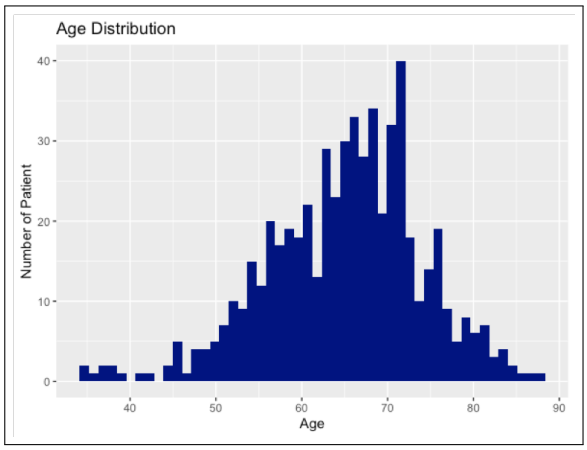

Figure 2: Age distribution of patients in our study

Figure 1 shows us the total number of patients experiencing a particular number of OFF periods in a week. Figure 2 shows us that most of the range of age of patients included in our study.

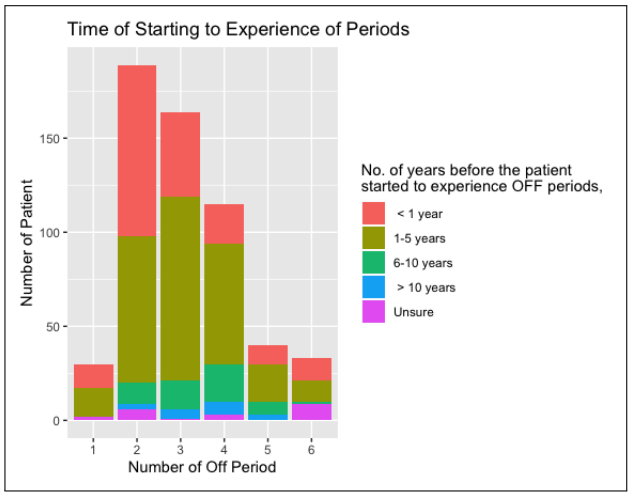

Figure 3: Time of Starting Experience of Periods

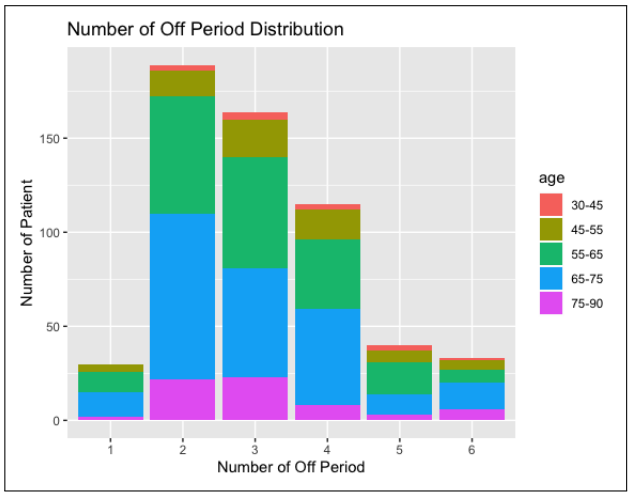

Figure 4: Number of Off Period Distribution

Figure 3 shows that most of the patients included in our dataset started to experience OFF periods 1 year before the data was collected.
Figure 4 shows us that most of the patients experience 2 to 3 OFF periods in a week. The age group that faces more OFF periods is 65 to 75 years.

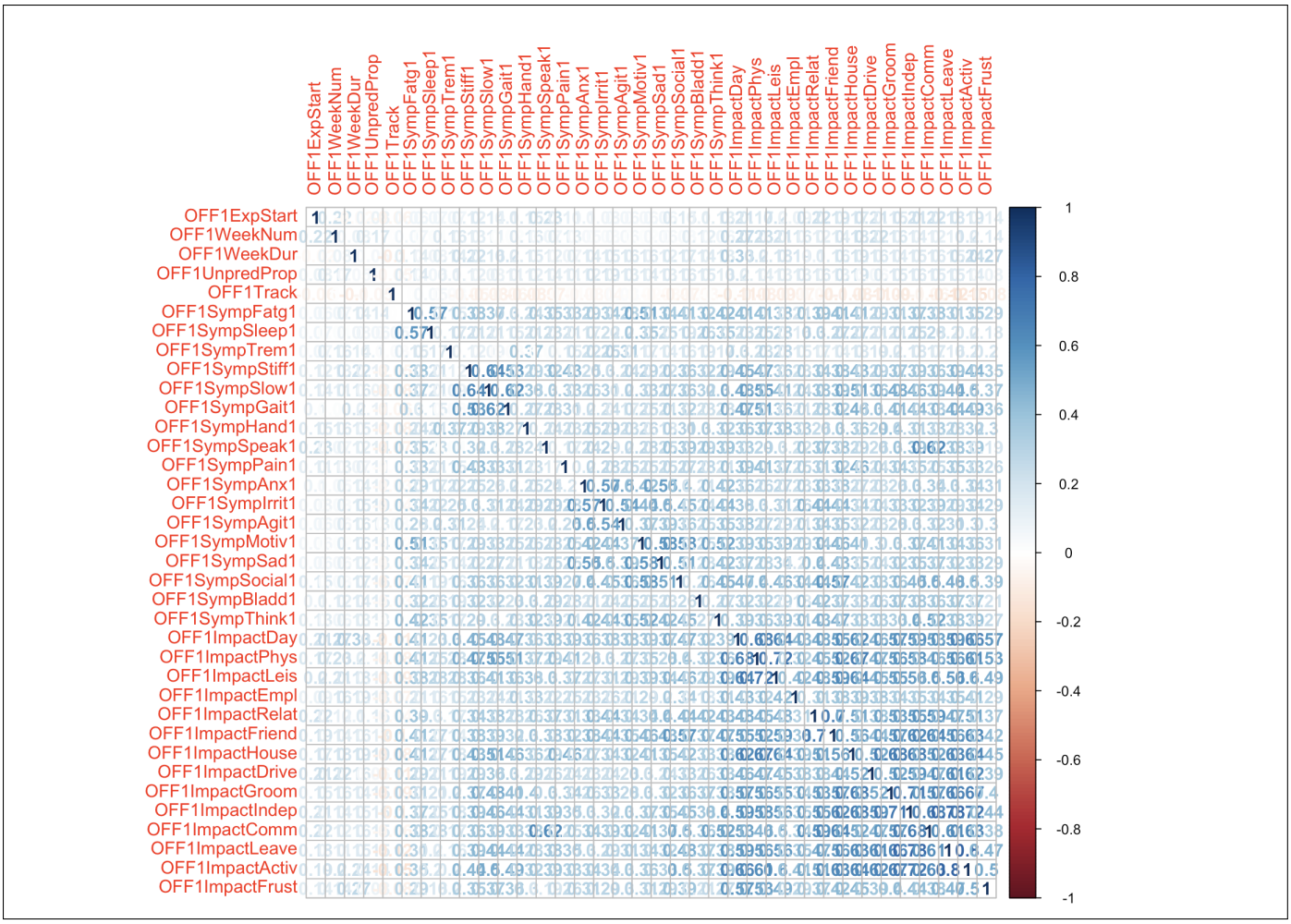

Figure 5: Correlation plot

Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. The right-side scale shows red and blue color gradients which represent negative and positive correlations respectively and the gradience represents magnitude. The circles are filled clockwise for positive values and anti-clockwise for negative values.

4 Methodology

We use supervised learning methods to develop various predictive models to predict the number of OFF periods for PD patients. Classification and Regression are two methods of supervised learning. The final data set obtained is split into training and test data sets. The training data set has 80 percent values from our final data set selected randomly. The rest 20 percent values form the test data set. In order to fix the values that go in training and test data sets, we use the set.seed() function. We train the training data set with a range of predictive models to predict our said response. The trained model is then tested on our test data set and model accuracy is assessed. Each of the predictive models used are briefly described below. We have used [9] for the content used to describe the predictive models. Additionally, we have referred [2] [1] for the formulas and describing algorithms in this section.

4.1 Multiple Linear Regression Model

Multiple Linear Regression establishes a relationship between our dependent variable and the independent variables using a best fit line (Regression line). A model that might describe this relationship is:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \epsilon$$

where,
Y_i is the dependent variable (response)
β_0 defines the intercept of plane,
$\beta_1, \beta_2, ...,\beta_i$are the regression coefficients. It measures the expected change in Y for a unit change in variable Xi when other variables are kept constant. The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. Design and Analysis of Experiments by Douglas C. Montgomery provides an easy description of the method of least squares. It estimates β_0, β_1, ..., β_i using the values that minimize:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Figure 6: Residual least squares equation; ref: ISLR

There are many alternatives to using least squares. We will discuss Subset selection and shrinkage further in this paper.

4.2 Multivariate Adaptive Regression Splines

In Multivariate adaptive regression splines (MARS), we start with the simplest model involving only the constant basis function i.e. increase its complexity by adding basis functions until a defined maximum level of complexity has been reached. Similar to backward subset selection, we Then we remove the least significant basis functions from the model. These functions are the ones which on removal will lead to minimum reduction in least squares [22] It can be represented as:

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

where,
B_i (x) is a function in collection of basis function or a product of 2 or more functions.

4.3 Best Subset Selection

Subset selection is the process that involves identifying a subset of predictors that we believe are more related to the response. Best subset selection process involves fitting a separate least squares regression for each

possible combination of the predictors. It starts with fitting 'p' models that contain just 1 predictor, the models with all combinations of 2 predictors and so on. The goal is to identify the best among all these models. Since the process of selecting best set of predictors from 2^p models is broken down into 2 stages:
1. A null model with 0 predictors which predicts sample mean for each observation.
2. Selecting the best model among all possible combinations of a total 'p' predictors
3. Selecting the best model among all the models tried using cross validated prediction error, Cp (AIC), BIC, or adjusted R^2.

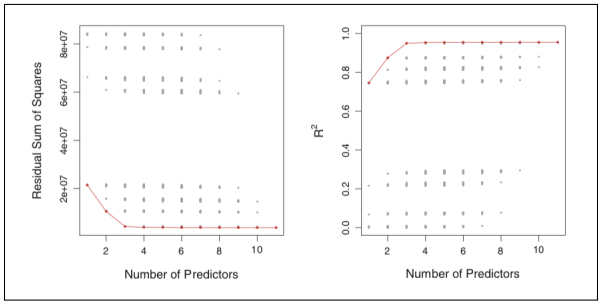

Figure 7: Sample plot that explains best number of predictors to be selected based in RSS and R^2; ref: ISLR

4.4 Forward Stepwise Selection

In best subset selection, we consider all 2^p possible models containing subsets of p predictors, forward stepwise selects a smaller set of models. It starts with null model and adds predictors to the model, one at a time until all predictors are in the model. The total number of models into consideration are substantially low in forward stepwise selection. It considers one null model, along with p-k models in the K^{th} iteration, where 'p' is the total number of predictors. Similar to best subset, the best model among all the models tried is selected using cross validated prediction error, Cp (AIC), BIC, or adjusted R^2.

4.5 Backward Stepwise Selection

Backward stepwise selection begins with the full least squares model containing all 'p' predictors, and then iteratively removes the least useful predictor, one-at-a-time. Similar to forward stepwise selection, backward selects from a total of $\frac{1+p(p+1)}{2}$ models. The best model among all the models tried is selected using cross validated prediction error, Cp (AIC), BIC, or adjusted R^2.

4.6 Ridge Regression

Shrinkage is another alternative to least squares. It involves using all 'p' predictors. The estimated coefficients are regularized to zero, which reduces variance. Depending on the type of shrinkage, some of the coefficients can be estimated to be equal to zero and thus these methods can also perform variable selection. The Ridge regression coefficient estimates are the ones that minimize the below equation:

$$\sum_{i=1}^{M} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} \omega_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} \omega_j^2$$

where,
$\lambda >= 0$ is a tuning parameter. Similar to least squares method, Ridge regression coefficients are those that fit the data well, by minimizing RSS. The shrinkage penalty $\lambda \sum_{j=0}^{p} \omega_j^2$ is small when the coefficients are close to zero, thus having a shrinking effect on coefficients. λ is the tuning parameter that is adjusted as per our target to get either a better fit or the size of coefficients.

4.7 Lasso Regression

Ridge regression has a final model that contains all 'p' predictors instead of selecting a subset. Lasso Regression allows us to overcome this disadvantage. The Lasso regression coefficient estimates are the ones that minimize the below equation:

$$\sum_{i=1}^{M} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} \omega_j \times x_{ij} \right)^2 + \lambda \sum_{j=0}^{p} |\omega_j|$$

Lasso uses a penalty that forces some of the coefficients to be exactly equal to zero when λ value is sufficiently large. Lasso is thus similar to subset selection in a way that it also uses a subset of predictors. Moreover, selecting the optimal value of λ is essential and cross validation is the method of choosing this value.

4.8 Generalized Additive Model (GAM)

As discussed earlier, Generalized linear models capture the linear relationship between the dependent and independent variables. We replace the linear function with a smooth non-linear function.
In other words, we replace β_i (X_{ij}) with $f_j(X_{ij})$.
It allows to retain the additive structure of Linear models while allowing flexible non-linearities in variables.
The GAM equation is as below:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j (x_{ij}) + \varepsilon_i$$

The models discussed earlier are Parametric. They are easy to interpret but the final model does not necessarily match the true form of function. It first assumes the shape of 'Y' and then use the training data to fit. Non-parametric methods on the other side are difficult to interpret. It avoids unrealistic assumptions and potential to accurately fit a wider range and shapes of function. However, these models are more susceptible to over-fitting. We discuss few non-parametric methods below:

4.9 Regression Trees

Tree based models basically segment the predictor space into a number of simpler rectangular or box shaped regions for easier interpretation. We make the predictions for observations in each region which is the mean of response values in that particular region. The aim is to find a region that minimizes the below equation:

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - y_{\hat{Rj}})^2$$

where, $y_{\hat{Rj}}$ is the mean response of observations in j^{th} box.
It begins at the top of the tree and then successively splits the predictor space (training) and each split is indicated via two new branches further down on the tree. The best split is made at each step without considering the future trees that will be built. By best split, we mean the one that will give us minimum RSS value. The process of splitting into trees in continued till we put a stopping criterion. We predict the response for test observations using the mean of training observations in the region to which the test observations belong. This process might over-fit the data leading to bad predictions. Smaller trees might lead to lower variance but there is a bias trade-off.
A strategy to overcome over-fitting is pruning where we grow a huge tree and prune it to get a smaller tree with less splits. A tuning parameter α controls a trade-off between the subtree's complexity (tree size) and its fit to the training data and the value of α is determined by cross-validation.

4.10 Bagging

It is a general procedure for reducing the variance of a statistical learning method. Number of decision trees are build on bootstrapped training samples. Here a random sample of m predictors are chosen as split candidates from a full set of p predictors. In this method, bootstrapping is done by taking repeated

samples from the training set. Here, B different bootstrapped training data-sets are generated then, training the method on the b^{th} bootstrapped training order in order to get the $\widehat{f}^{*b}(x)$ and finally averaging all the predictions , to get

$$\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{*b}(x)$$

4.11 Random Forest

Random forest decorrelates the trees which leads to reduced variance when we average the trees. Similar to bagging, decision trees are built on bootstrapped training samples. However, when building these trees, a random set of predictors are chosen from the complete set 'p' and the split only uses these random subset of predictors. Each split will have a new subset of 'm' predictors selected randomly. Typically, the value of 'm' in $\sqrt{p}$. Using this relation leads to reduction in best test error and OOB (out of bag) error over bagging. It is up to us to choose the value of 'm'.

4.12 Boosting

Boosting is similar to bagging in a way that it creates multiple copies of Training data set by bootstrapping and a decision tree is fit to each copy. The final predictive model is a combination of all these trees. The major difference in Boosting is that the trees are built sequentially i.e. every tree is built based on information from previous trees. Boosting avoids the issue of overfitting since we are not fitting a huge tree to the entire data set and it learns sequentially. The tuning parameter are number of trees. We use cross validation to determine the number of trees. The shrinking parameter λ determines the rate at which boosting learns. The number of splits in each tree controls the complexity of boosted ensemble. A single split tree works well and provides an additive model.

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x). \quad (8.10)$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \quad (8.11)$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \quad (8.12)$$

Figure 8: Boosting algorithm for Regression trees; ref: ISLR

5 Comparison and Results

We consider Mean Square Error (MSE) as the baseline for comparison since it tells us how close a regression line is to the set of points. The distances from points to regression line is the error. We square it cancel out the negative signs values i.e points on either side of regression line. While checking the MSE values, our aim to minimize MSE since we want to minimize bias and variance. The below mentioned tables compares the final model MSE values and Training MSE values for each model.

Predictive Model	Train MSE
Multiple Linear Regression model	1.149
MARS	1.121
Best subset	1.112
Forward stepwise	1.112
Backward stepwise	1.112
Ridge	1.375
Lasso	1.301
GAM	1.017
Regression trees	1.268
Bagging	0.389
Random Forest	0.255
Boosting	0.298
Null Model	1.410

Table 4: Comparison of Training MSE

Predictive Model	Test MSE
Multiple Linear Regression model	1.712
MARS	1.908
Best subset	1.939
Forward stepwise	1.939
Backward stepwise	1.939
Ridge	2.017
Lasso	1.947
GAM	2.014
Regression trees	1.035
Bagging	1.035
Random Forest	1.123
Boosting	1.553
Null Model	2.033

Table 5: Comparison of Test MSE

Based on the above mentioned values, Bagging is the best fit predictive model for our data set since it has the least MSE value. There is almost 50 percent reduction in the MSE value as compared to Null model. Further, we analyze the important variables obtained from Bagging.

5.1 Important Variables

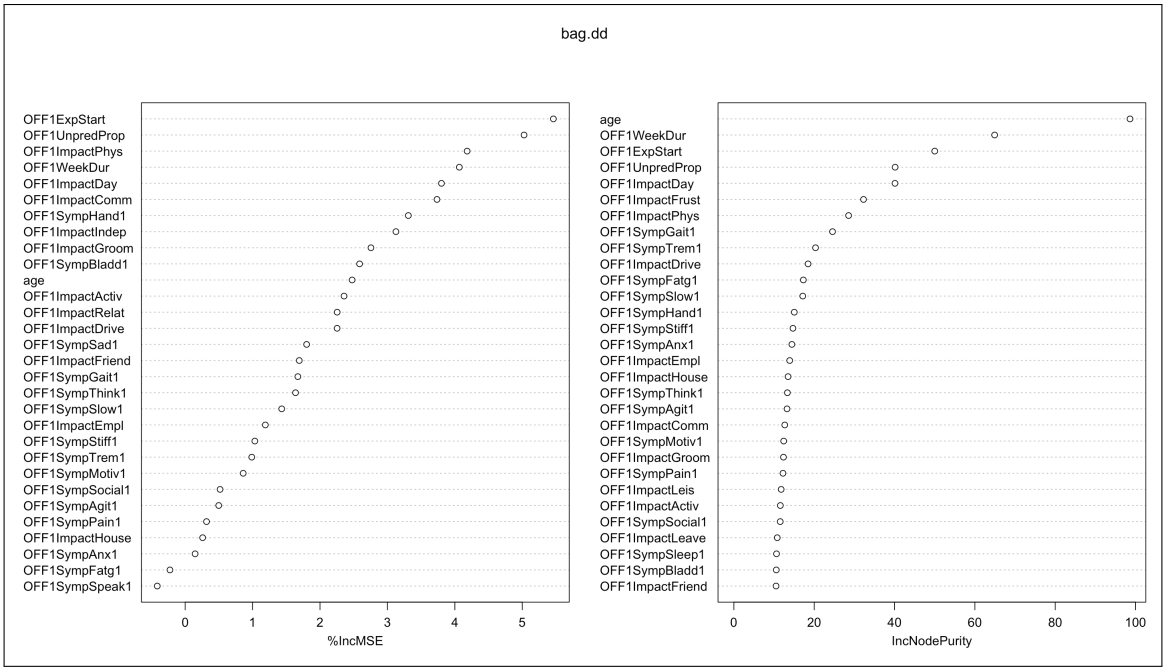

Figure 9: Important variables

In the bagging model, we start with 500 trees. After determining the optimal number of trees (170), the model is improved. Totally 37 variables are used in the model. The significant variables are observed from the left side of the Figure 7. The first time of experiencing off periods has the most effect on MSE. Moreover, proportion of the unpredictable off periods follows as an important variable. The impact on physical activity

is another factor for increasing MSE value. In the following section, it will be considered that each predictors individually dependency on predicted value.

5.2 Partial Dependence Plots

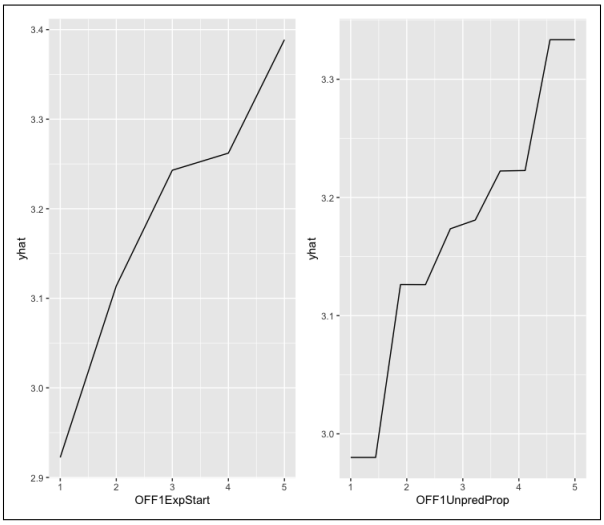

Figure 10: The left plot implies that as the number of years ago the patients started to experience OFF periods increases, the prediction of number of OFF periods of a patient in a week increases. For the right plot, as the proportion of patient's OFF periods increases at unpredictable time, the prediction of the number of OFF periods of a patient increases.

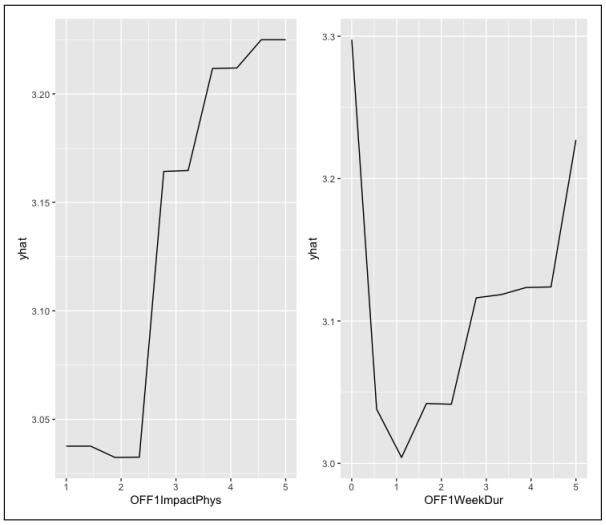

Figure 11: The left plot tell us that as the patient's physical activity increases the prediction of the number of OFF periods of a patient increases. For right plot, the prediction of the number of OFF periods decreases as the duration of each OFF period of a patient (average) increases till a certain point, after the point it increases

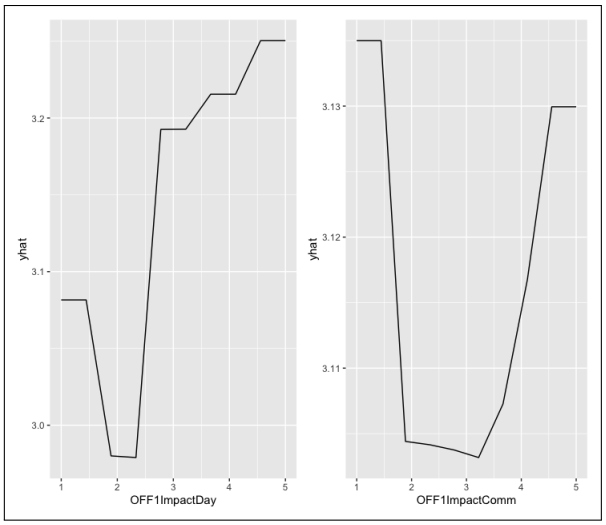

Figure 12: In left plot, the prediction of the number of OFF periods decreases as the impact of the OFF periods increases at a certain point, after that it start increasing. Similarly for the right plot, the prediction of the number of OFF periods decreases as the communication of the patient increases at a certain point, after that point it start increasing.

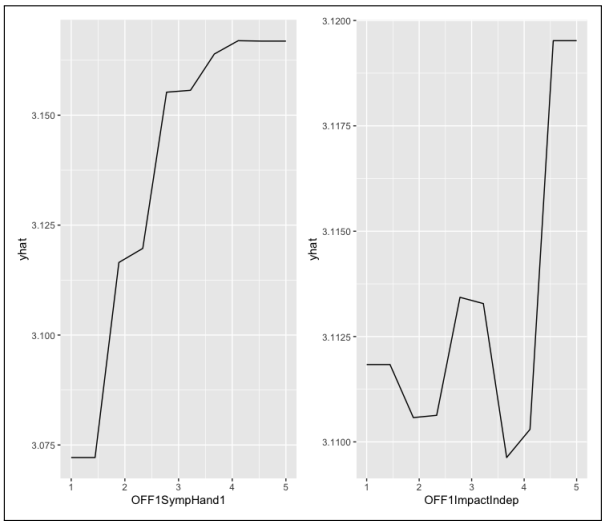

Figure 13: The left plot tell us that as the patient's difficulty increases for hand coordination, the prediction of the number of OFF periods increases. For the right plot, as the independence of the patient increases the prediction decreases at first and then it increases for a certain point and again it decreases at a certain point and then again it increases.

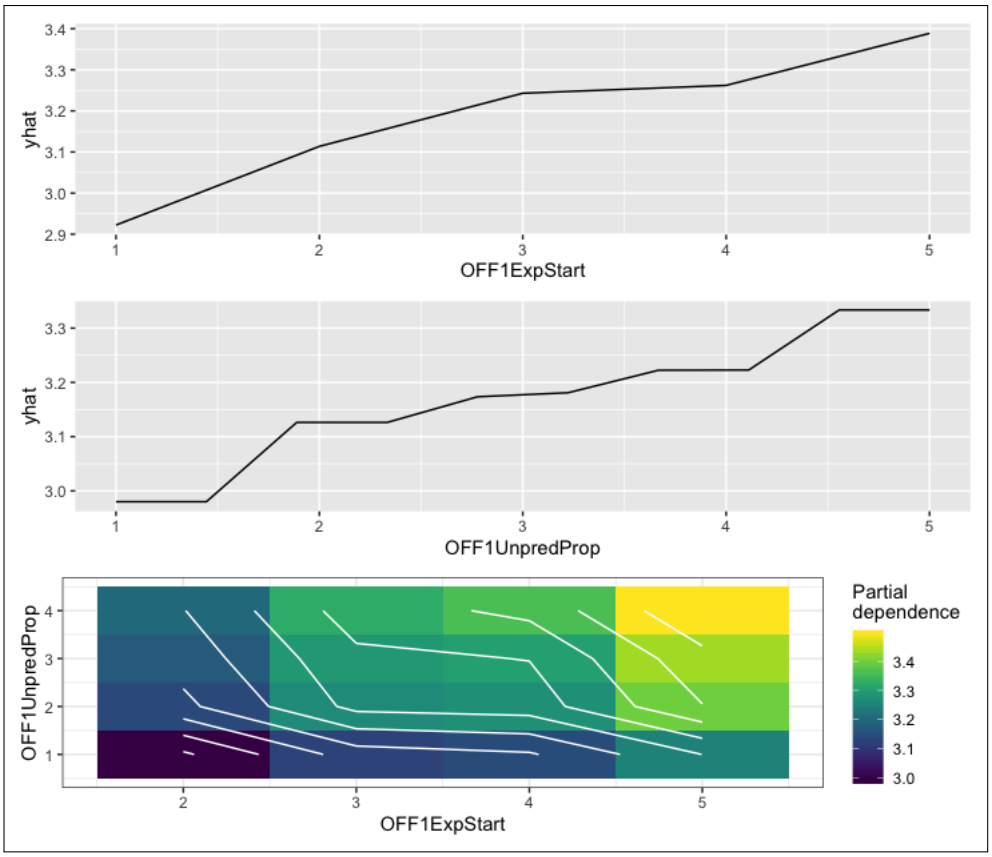

Figure 14: Total effect on predicted values by the most significant variables

6 Conclusion

In this study, the aim was predicting the number of off periods in a week for patients. We considered the symptoms and it's severity each patient experiences and the magnitude of impact it has on their daliy activities. We used parametric, semi-parametric and Non-parametric predictive models to find the best fit for our data set. The predictive model with least MSE value, our defining parameter, is the best fit. The difference in Test MSE and Train MSE for bagging obtained is acceptable and shows us that model was not overfit. In the end, we determined the significant variables for predicting OFF periods. Many studies suggest that age is an important parameter for Parkinson's Disease [13]. However, in this study, we found that age is not that significant. One of the reasons for this we believe is that PD commonly arises in old age. Since the disease is progressive, the symptoms go on becoming severe with age. Most of the patients that showed severe symptoms were above the age of 60 years. The number of years before the patient starts experiencing OFF periods is the most significant factor. Apart from this, the patients should note how difficult it is for them to Drive, do regular scheduled activities, be independent and communicate. They should consult their medical advisor if they frequently feel depressed/sad, have difficulty in thinking, have difficulty in hand coordination and difficulty in Bladder control.

7 Future Scope

In this study, we developed a model to predict the number of OFF periods the patient will experience in a week, given the severity of symptoms. In future, with literature survey, we can study the kind of medication prescribed for PD and then consequently, try to match the proper medication and it's required doses based on the severity. [17] provides a good read for importance of medications in Parkinson's Disease treatment.

References

[1] Analytics Vidhya, howpublished = https://www.analyticsvidhya.com/myfeed/?utm-source=blogutm-medium=top-icon Accessed: 2019-11-15.

[2] Towards data science, howpublished = https://towardsdatascience.com/@tonester524. Accessed: 2019-11-15.

[3] UPDRS Development Committee et al. Recent developments in parkinson's disease. 1987.

[4] Diederik van de Beek, Ben Schmand, Jan de Gans, Martijn Weisfelt, Heleen Vaessen, Jacob Dankert, and Marinus Vermeulen. Cognitive impairment in adults with good recovery after bacterial meningitis. *The Journal of infectious diseases*, 186(7):1047–1052, 2002.

[5] Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience*, 5(1):s13742–016, 2016.

[6] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

[7] Ram Deepak Gottapu and Cihan H Dagli. Analysis of parkinson's disease data. *Procedia Computer Science*, 140:334–341, 2018.

[8] Margaret M Hoehn and Melvin D Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427, 1967.

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[10] Nicolas Khoury, Ferhat Attal, Yacine Amirat, Latifa Oukhellou, and Samer Mohammed. Data-driven based approach to aid parkinson's disease diagnosis. *Sensors*, 19(2):242, 2019.

[11] Muriel Deutsch Lezak, Diane B Howieson, David W Loring, Jill S Fischer, et al. *Neuropsychological assessment*. Oxford University Press, USA, 2004.

[12] Indrajit Mandal and N Sairam. New machine-learning algorithms for prediction of parkinson's disease. *International Journal of Systems Science*, 45(3):647–666, 2014.

[13] Amy Reeve, Eve Simcox, and Doug Turnbull. Ageing and parkinson's disease: why is advancing age the biggest risk factor? *Ageing research reviews*, 14:19–30, 2014.

[14] Robert S Schwab. Projection technique for evaluating surgery in parkinson's disease. In *Third symposium on Parkinson's disease*, pages 152–157. E&S Livingstone, 1969.

[15] Matthew B Stern, Anthony Lang, and Werner Poewe. Toward a redefinition of parkinson's disease. *Movement disorders*, 27(1):54–60, 2012.

[16] Anita Valmarska, Dragana Miljkovic, Nada Lavrač, and Marko Robnik-Šikonja. Analysis of medications change in parkinson's disease progression data. *Journal of Intelligent Information Systems*, 51(2):301–337, 2018.

[17] J Gert van Dijk, J Haan, Koos Zwinderman, Berry Kremer, BJ Van Hilten, and RA Roos. Autonomic nervous system dysfunction in parkinson's disease: relationships with age, medication, duration, and severity. *Journal of Neurology, Neurosurgery & Psychiatry*, 56(10):1090–1095, 1993.