Marquette University

COSC 3570 Introduction to Data Science

Fall 2023

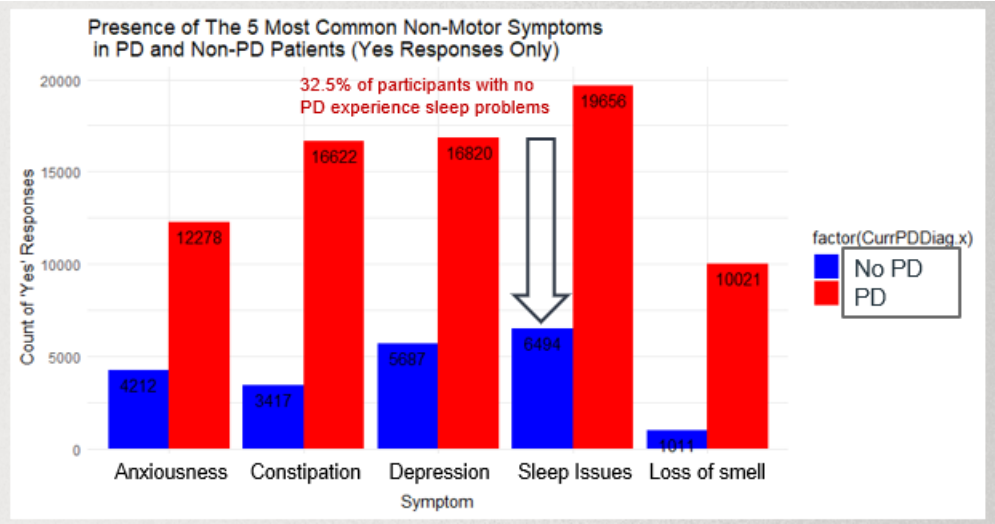
Nathan Estrin

Risk and Rate of Progression Detection Model for Parkinson's Disease

Introduction:

Every 6 minutes, someone in the United States is diagnosed with Parkinson's Disease , translating to an alarming 90,000 new cases each year. By the year 2030, it is anticipated that 1.2 million individuals in the US will carry this diagnosis (Parkinson's Foundation). The ramifications of Parkinson's extend beyond the personal toll on affected individuals, reaching into the economic sphere with an annual expenditure of approximately $52 billion on treatment. This financial burden is exacerbated by the lack of early intervention and the challenges associated with accurately diagnosing Parkinson's disease in its early stages.

To shed light on the urgency of my endeavor, I conducted an exploratory visualization study focusing on participants experiencing the top 5 early onset nonmotor symptoms of Parkinson's. This visualization includes responses from both diagnosed and undiagnosed participants experiencing the top five most common early onset non-motor symptoms. What emerged as a striking revelation was the prevalence of sleep disturbances, one of the key nonmotor symptoms associated with Parkinson's. Notably, 32.5% of undiagnosed participants reported experiencing sleep disturbances like those of undiagnosed individuals.



This observation raises a crucial concern: the potential underdiagnosis of Parkinson's disease (PD). The existence of undiagnosed cases implies a substantial gap in our current understanding of the disease's prevalence, hindering effective public health responses and medical interventions. In response to this revelation, the research at hand aims to contribute to bridging the diagnostic gap and, by extension, alleviating the socioeconomic burden associated with Parkinson's disease.

Parkinson's disease is the second most common chronic and progressive neurodegenerative disorder that is characterized by the degeneration of dopaminergic neurons in a region of the brain called the substantia nigra. This neuronal loss leads to a deficiency of dopamine, a neurotransmitter that is crucial for regulating movement. The exact cause of this neurodegeneration remains elusive, although a combination of genetic and environmental factors is believed to contribute.

Data Collection:

For this endeavor, a rich foundation was established through the utilization of data from a cohort of 50,000 participants enrolled in an ongoing study facilitated by the Michael J. Fox Foundation. The dataset comprises individuals that are diagnosed and not diagnosed with Parkinson's. Detailed demographic information, including age, gender, and ethnic backgrounds were collected to facilitate a nuanced examination of Parkinson's disease within diverse populations. Clinical characteristics such as age of onset, presence of non-motor and motor symptoms, and response to treatment were also considered. This wealth of information enabled a comprehensive exploration of potential risk factors and early indicators of disease progression.

	fox_insight_id	toxicant	pesticide	caffeine	alcohol	genetics	head	smoking	AgeofDiag
1	FOX_000014	NA	NA	NA	NA	NA	NA	NA	NA
2	FOX_000076	NA	NA	NA	NA	NA	NA	NA	81.3
3	FOX_000087	NA	NA	NA	NA	1	NA	NA	54.4
4	FOX_000126	1	NA	NA	NA	NA	1	NA	NA
5	FOX_000126	NA	3	1	1	NA	NA	1	NA
6	FOX_000126	NA	NA	NA	NA	NA	NA	NA	48.6

A notable limitation that came with the data is the presence of incomplete observations and multiple observations of the same participant with values for different columns in different rows. The reasons for missing data may vary, including participant non-compliance, dropout from the study, or technical issues during data collection. The data quality can be compromised when dealing with multiple observations for the same participant. It raises questions about the consistency and accuracy of the information collected. This issue introduces a challenge to

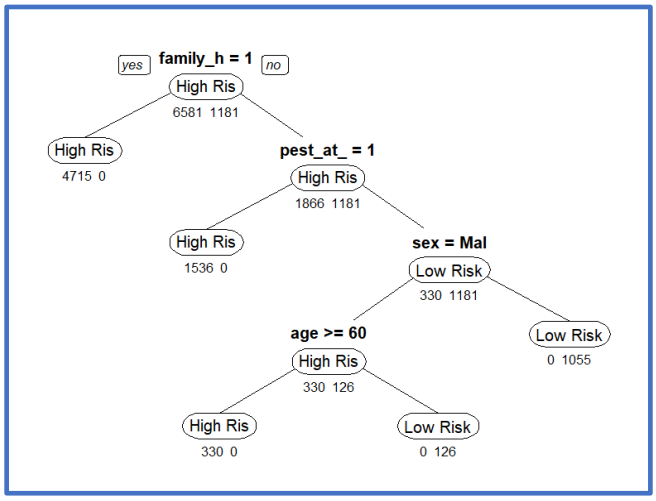
ensure the completeness and representativeness of the analyses conducted. Furthermore, these limitations may introduce bias and affects the generalizability of findings to the broader population. Above is a picture of a part of one of my datasets and provides an example of what was mentioned.

Model Development:

The development of my model involved a multifaceted approach, utilizing various statistical methods to address different aspects of the disease trajectory.

CART Analysis:

To ascertain the risk of developing Parkinson's disease within the dataset, a Classification and Regression Tree (CART) analysis was conducted using R, incorporating powerful functions from the Tidyverse, such as the mutate function and piping techniques. The mutate function facilitated the creation of a new variable called "risk category," a pivotal step in our analysis. Drawing on subject matter expertise, participants were classified as being at high risk of developing Parkinson's if they met specific criteria. The criteria included having a family history of Parkinson's disease, being a male older than 60, or having been exposed to high amounts of pesticides at home. These informed classifications were crucial for creating a nuanced risk assessment model, allowing for the identification of key predictors within the dataset. The use of tidyverse techniques not only enhanced the efficiency of data manipulation but also contributed to the transparency and reproducibility of the analysis.



The classification tree analysis yielded insightful results, unraveling the intricacies of Parkinson’s disease risk assessment within the dataset.

The root node, representing the variable with the most significant predictive power or the variable that created the purest subset was if one person has a family history of Parkinson’s

Disease. These individuals were predominantly classified as being at high risk. This highlights the substantial influence of genetic factors in determining the risk of developing PD.

For those without a family history, the presence of pesticides at home emerged as the next most significant predictor. This finding underscores the potential role of environmental factors.

In the absence of both family history of PD and pesticide exposure, the tree revealed that sex becomes a critical determinant of risk assessment. Females and males are considered at low risk. However, an exception was identified for males above the age of 60, placing them at high risk of developing Parkinson's. "There are clear sex-related differences in epidemiological and clinical features of the disease: PD affects men twice more often than women, but women have a higher mortality rate and faster progression of the disease (Cerri). " Existing research has even proved that men are at higher risk of developing Parkinson's. This nuanced finding emphasizes the interaction between age and gender as additional risk factors in specific subgroups.

Multiple Linear Regression:

Another alternative to early detection of PD is predicting the age at diagnosis. I constructed multiple MLRs using commons risk factors such as excessive caffeine intake, excessive alcohol intake, excessive smoking, head injuries, exposure to pesticides and exposure to toxicants as my predictors. Unfortunately, due to missing values in the age of diagnosis variable, imputation was performed using the median of existing values.

	Model 1	Model 2
Intercept	57.48694 (2e-16)**	58.5977 (2e-16) **
toxicant	-1.12470 (0.16487)	--
pesticide	0.53439 (0.68679)	--
caffeine	3.29656 (9.4e -05) **	3.2799(7.53e-05)**
alcohol	0.06869 (0.93771)	--
smoking	2.14372 (0.00459) **	2.0711(0.00533)**
head	-2.04027(0.00597)**	-2.0981(0.0047)**
genetics	1.50990 (0.08962)	--

Se	9.431	9.438
R^2	0.05093	0.04445
Adjusted R^2	0.04203	0.04062
F-statistic	5.719 (1.848e-06)	11.63 (1.868e-07)

Two MLR models were developed. The first included all the variables, while the second was refined based on the significance of variables determined by their p-values. Interestingly, the intercept in the second model indicated a baseline age of Parkinson's disease diagnosis at approximately 58 years when all other variables were held constant.

I discovered that caffeine consumption emerged as a significant predictor, associated with an increase in the age of diagnosis by about 3.3 years when other variables were held constant. This suggests that higher caffeine intake might correlated with a later onset of PD. Similarly, smoking was associated with an increase in the age of diagnosis by about 2 years, indicating that those who smoke tend to be diagnosed at an older age compared to non-smokers. In parallel to my results, existing research has probed into the protective effects of both caffeine and smoking in the context of PD. "Like caffeine, nicotine has been found to reduce MPTP-induced dopaminergic toxicity in animal models of Parkinson's disease. One mechanism underlying this protective action may be its ability to increase the expression of neurotrophic factors that are known to promote survival of dopaminergic neurons. But tobacco contains numerous other chemicals whose influence on biological processes may play a part. Smoking causes a reduction in activity of monoamine oxidase A and B, for example, which might protect against neuronal damage by inhibiting the enzymatic oxidation of dopamine (Gale)."

In contrast, head injuries appeared to decrease the age of diagnosis by about 2 years when all other variables were held constant, suggesting that neurological trauma may lead to an earlier onset of Parkinson's disease.

Despite the statistical significance of Model 2, it only explained about 4% of the variance in the age of diagnosis, indicating a low predictive power. This low r-squared value suggest that unaccounted factors might influence the age of Parkinson's disease diagnosis, reflecting the complexity of the disease and the challenge in predicting its onset age solely based on common risk factors. This limitation underscored the need for further exploration and consideration of additional factors contributing to the age of PD diagnosis.

Logistic Regression Model:

Transitioning from the predictive models detecting PD and its onset age, the logical nest step in the continuum of patient care involves understanding whether the disease is progressing. To address this critical aspect, a logistic regression model was developed, leveraging a range of clinical factors to determine the progression status of Parkinson's disease in diagnosed individuals.

In crafting the model, a binary variable for disease progression was created, drawing upon clinically relevant thresholds for various indicators. A value of 1 indicated disease progression, while a 0 signified the absence of progression.

	Logistic Model
Intercept	-1.88975 (2e-16)**
MoveSpeech	0.18603(4.43e-09)**
MoveWrite	0.01956 (0.466)
MoveTremor	0.22345 (2.76e-11)**
Mobility	0.06991(0.110)
Care	-0.06041(0.272)
Active	0.36577 (4.71e-14)**
Pain	0.21506 (9.11e-09)**
Anxious	0.47293 (2e-16)**

Notably, the investigation unveiled that several common indicators of disease progression in Parkinson's patients were associated with an increase in the log odds, with a standout factor being the level of anxiety and depression. A one-unit increase or moving up one level in this aspect was found to be significantly correlated with a 0.473 increase in the log odds of disease progression.

The evaluation of the logistic regression model's performance is pivotal in gauging its accuracy. The model demonstrated an accuracy rate of 73.85%, indicating that in nearly three-quarters of the cases, it accurately predicted the progression status of the disease. This percentage was derived by comparing the model's predictions with the observed outcomes in the dataset, using the pred() function and expressing the result as a percentage. This robust accuracy underscores the reliability of the logistic regression model in effectively discerning the progression status of PD based on relevant clinical presentations.

Synthesis:

The culmination of the developed risk and progression detection models for Parkinson's disease not only advances our understanding of the disease but also holds transformative potential for the field. This synthesis underscores the anticipated impact on patient outcomes, the reduction of healthcare costs, and the propelling force these models offer for future investigations.

Improving Patient Outcomes:

One of the primary objectives of this research is to significantly improve patient outcomes in the realm of Parkinson's disease. By accurately predicting the risk of disease development and progression, these models equip clinicians with valuable tools for early intervention and personalized treatment strategies. Patients identified at high risk can benefit from proactive measures, potentially delaying or mitigating the onset of Parkinson's symptoms. For those already diagnosed, the logistic regression model provides insights into the likelihood of disease progression, enabling tailored treatment plans. Ultimately, these models contribute to enhanced patient care by facilitating more precise interventions at critical junctures in the disease trajectory.

Reducing the Need for Intensive Treatment:

The implementation of predictive models in Parkinson's disease care has the potential to reduce the need for intensive treatments, particularly for those identified as at high risk or experiencing disease progression. Early intervention strategies based on the risk assessment model may mitigate the severity of symptoms, potentially minimizing the reliance on aggressive therapeutic approaches. Moreover, the logistic regression model aids in identifying individuals with a lower likelihood of disease progression, allowing for a more nuanced and conservative treatment approach. By tailoring interventions to individual risk profiles, the models aim to optimize treatment efficacy while minimizing the burden of intensive interventions on patients.

Reducing Healthcare Costs:

The economic impact of Parkinson's disease on healthcare systems is substantial. By providing tools for early intervention and personalized care, the developed models have the potential to significantly reduce healthcare costs associated with the management of Parkinson's disease. Early identification of at-risk individuals allows for targeted interventions that may be less resource-intensive compared to managing advanced stages of the disease. Moreover, the

logistic regression model's ability to predict disease progression aids in resource allocation, directing intensive treatments toward those who stand to benefit the most. This targeted and personalized approach has the potential to alleviate the economic burden associated with Parkinson's disease management.

Setting the Stage for Future Investigations:

Beyond the immediate clinical applications, the developed models lay a robust foundation for future investigations in the field of Parkinson's disease research. The nuanced risk assessment model provides insights into complex interactions among genetic, environmental, and demographic factors. Researchers can build upon this framework to explore additional risk factors, refine existing models, and delve deeper into the intricacies of Parkinson's disease etiology. Furthermore, the logistic regression model, with its focus on disease progression, sets the stage for investigations into novel therapeutic targets and interventions. By identifying key predictors of progression, future research can explore interventions that specifically target these factors, opening avenues for innovative and targeted treatment strategies.

Works Cited:

Cerri S, Mus L, Blandini F. Parkinson's Disease in Women and Men: What's the Difference? J Parkinsons Dis. 2019;9(3):501-515. doi: 10.3233/JPD-191683. PMID: 31282427; PMCID: PMC6700650.

Foundation, Parkinson's. "Prevalence & Incidence." *Parkinson's Foundation*, www.parkinson.org/understanding-parkinsons/statistics/prevalence-incidence. Accessed 10 Dec. 2023.

Gale C, Martyn C. Tobacco, coffee, and Parkinson's disease. BMJ. 2003 Mar 15;326(7389):561-2. doi: 10.1136/bmj.326.7389.561. Erratum in: BMJ. 2003 Mar 22;326(7390):614. Gale, Chris [corrected to Gale, Catharine]. PMID: 12637374; PMCID: PMC1125458.