**Final Report**

Sankhepo Ndhlovu

Southern New Hampshire University

MAT 430: Seminar in Applied Mathematics

Timothy Smith

August 13, 2023

1. **INTRODUCTION**
2. **Topic**

Parkinson’s disease (PD) is the most complex motor-related brain disease and the second most common neurodegenerative disorder, following dementia (Yang, F. et al., 2016). Affecting over 1% of the global population, PD is typically characterized by abnormally low levels of dopamine due to dead or impaired nerve cells in the basal ganglia of the brain (Hoseinipalangi, Z. et al., 2023). Symptoms include tremors and rigidity, which significantly impact cognitive function, emotional well-being, and overall quality of life. Parkinson’s commonly develops in the later stages of life, affecting individuals over the age of 60 (Yang, F. et al., 2016), although it can occur in very rare cases in individuals below 35 years old. The exact causes of PD remain unknown, but emerging studies suggest that social determinants may play a significant role in its prevalence. Therefore, this research paper aims to investigate the social determinants that contribute to the prevalence of PD.

My interests in this topic stem from personal experiences with inherent health disparities as well as a close connection to an individual diagnosed with Parkinson’s disease. Witnessing their decline through the gradual loss of cognitive and physical functions over the past years has motivated me to explore healthcare and research opportunities, with a goal of understanding such disease patterns and evaluating their connection to social determinants as well as socioeconomic status. This interest is supported by my proficiency in regression analysis that drives me to translate numbers into actionable insights that could potentially lead to better care, enhanced interventions, and ultimately, improved quality of life for individuals living with PD. Moreover, as a mathematics major nearing the conclusion of my undergraduate journey, my academic journey is built upon quantitative analysis, data interpretation, and statistical modeling. Therefore, these growing skills will prove beneficial in applying statistical methods and large datasets to medical and health-related research.

1. **Research Question**

This study aims to explore the relationship between social determinant and PD prevalence by examining the influence of various factors including age, race/ethnicity, gender, income, educational level, occupation, U.S. born or foreign-born status.

The medical community's research on PD predominantly focuses on its physical aspects; however, its etiology remains multifaceted and poorly understood. Present treatments primarily target symptomatic relief rather than halting disease progression (Yang, F., et al. (2016)). These gaps place a significant burden on the healthcare system and economy, with PD's annual economic impact in the U.S. reaching $52 billion (Yang, W., et al. (2020)). Moreover, individuals with Parkinson's face a twofold higher mortality rate compared to the general population (Wirdefeldt, K., et al. (2011)). My research seeks to bridge these gaps. Additionally, earlier studies have underscored the significance of factors such as age and gender in shaping health outcomes, including neurological disorder prevalence. Therefore, my question specifically seeks to address, “What effect does gender have on Parkinson’s disease prevalence?” Furthermore, I aim to explore, “Does this effect of gender on PD vary or depend on other determinants like age?”

1. **Information**

The dataset selected for this study was obtained from the Inter-university Consortium for Political and Social Research (ICPSR). It encompasses 139 observations and 822 variables, initially used for a study on “The Emergence and Evolution of Social Self-management of Parkinson’s Disease” within the greater Boston Metropolitan Area between the years 2013 and 2019. The variables chosen for analysis from this dataset were gender, age, education, and income.

1. **Assessment**

Based on the information I have gathered, my underlying assumption is that a strong positive correlation exists between certain determinants and PD prevalence. Having a range of variables related to Parkinson's disease and potential determinants, this dataset will help create an effective model to tackle my research question concerning the determinants effects on PD.

However, it's important to acknowledge the assumptions and limitations inherent within the dataset. Firstly, the dataset is area specifical to the greater metropolitan area of Boston, which might affect the generalizability of my findings, as they may not fully represent the entire Parkinson’s disease population. Next, the dataset's self-reported nature might introduce biases and inaccuracies in the variables' measurement, potentially impacting the model's precision. Additionally, certain determinants that influence Parkinson's prevalence, such as ethnicity and race have limited representation in this dataset. Finally, I recoded for variables for income, education levels, and age because they initially comprised numerous categories and relatively few observations, which could complicate the analysis.

1. **MODEL**
2. **Selection**

Logistic regression was selected for my research due to its compatibility with the categorical outcome variable of the research question (Parkinson's disease prevalence) and its ability to handle multiple predictor variables. Logistic regression is specifically designed to model binary outcomes.

Compared to other models, logistic regression is an extension of linear regression. However, while linear regression excels at predicting continuous outcomes, it falls short when faced with binary outcomes such as disease prevalence, as it could generate predictions beyond the range of 0 and 1.

Furthermore, logistic regression is used in healthcare research to explore the impact of one or more exposure variables on disease outcomes. Notably, previous studies, including one by Park, R. S et al. (2021), successfully utilized the robustness of logistic regression for investigating health outcome determinants. Additionally, a study by Schober, P., & Vetter, T. R. (2021), underscores the significance of logistic regression in situations where the outcome is dichotomous, as it is in my research.

However, despite its merits, logistic regression does possess limitations. Its assumption of linearity between the logit of the outcome and the independent variables may not always hold, given the fact that complex relations may exist. To assess this assumption, I will consider plotting the log odds against the predictors. Additionally, comparative tests such as likelihood ratio tests and model fit statistics like the AIC or the BIC, which measure the model's goodness of fit, will be performed to support the model choice.

1. **Creation**

The construction of a logistic regression model is important to understanding the connection between social determinants and Parkinson's disease prevalence.

The Formula of Insight is expressed as:

However, as this formula results in a slope that is not constant, neatly summarizing the relationship between the estimators and the outcome becomes nearly impossible (Mesa J. L. (2004). Therefore, a logit transformation will be used, which enables the modeling of log-odds of the outcome variable as linear functions of the independent variables. This ensures a constant slope which result in meaningful interpretations.

The logit transformation model is expressed as:

Where:

* is the natural logarithm.
* is the probability of the disease outcome.
* is the odds of disease outcome.
* is interpreted as “log - odds”.
* is the intercept.
* are the coefficients for each respective predictor.
* are the predictor variables.

Fitting my selected variables, the logistic regression model can be expressed as:

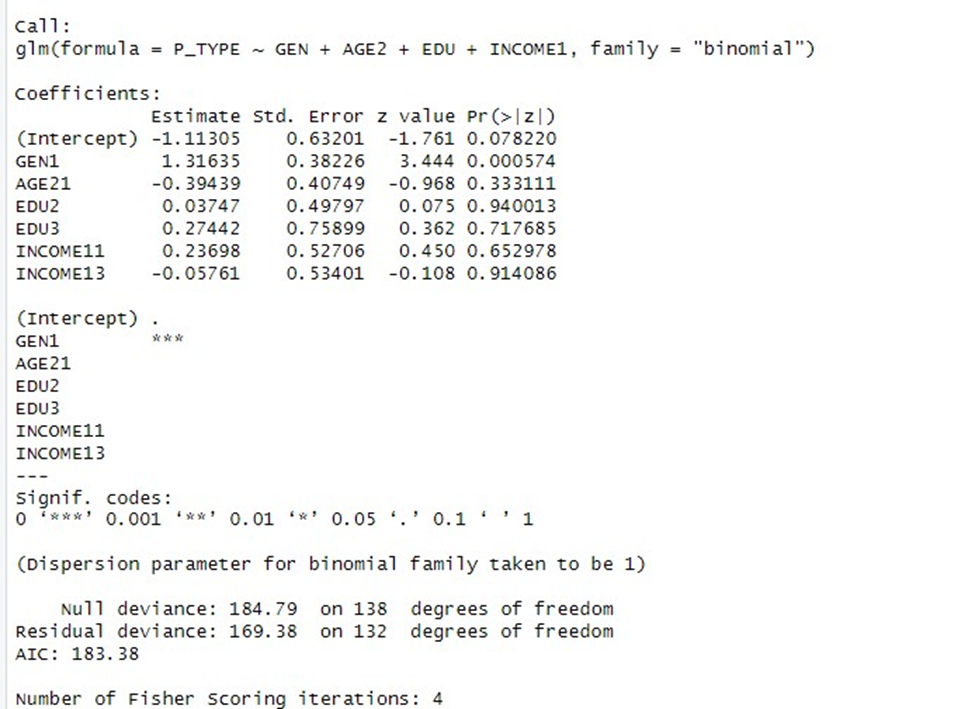
Where:

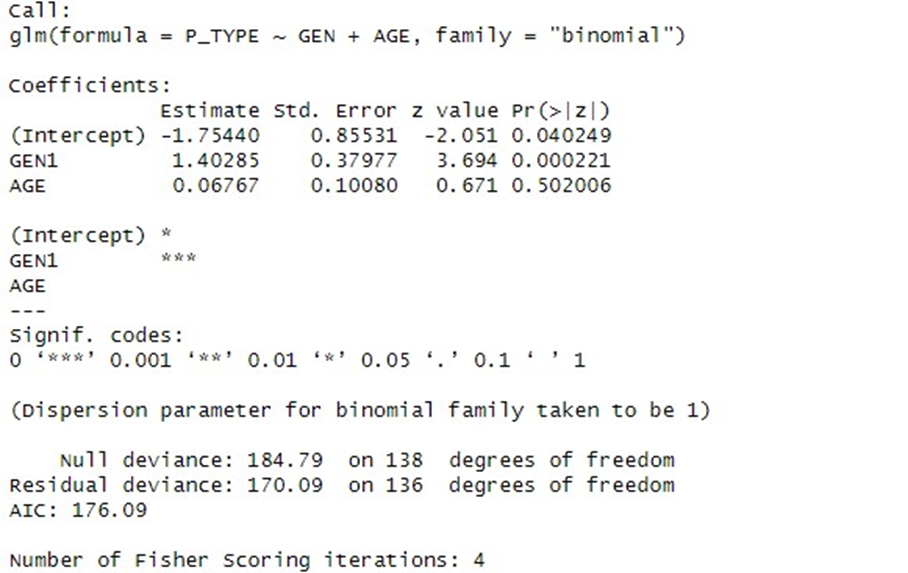
* + GEN (gender) is Male = 0, Female = 1
  + AGE, less than 65 years = 0, greater or equal to 65 years old = 1
  + EDU (education), some college or no degree = 1, Associates, bachelor’s, or master’s degree = 2, professional school or doctoral degree = 3
  + INCOME, less than $5000 = 0, between $5000 & $74,999 = 1, over $75,000 = 2.

R programming language was employed to build the model, which ensured precision in coefficient estimation and in handling calculations involved.

The initial model incorporated all the variables: gender, age, education, and income. Education and income were considered potential confounding factors because, conceptually speaking, they could both correlate with elements such as healthcare access, lifestyle choices, and environmental exposures that directly or indirectly impact Parkinson's disease prevalence. Consequently, to account for their potential confounding effects, they were included as covariates in this initial model.

However, for this model, the elevated AIC (183), increased residual deviance (169), and presence of only one significant estimator (GEN) indicate that the inclusion of income and education as covariates was not statistically significant. This suggests that the model might be better equipped to predict the outcome of Parkinson’s without their inclusion.



Therefore, a second model was fitted, excluding education and income. This model appeared somewhat promising, exhibiting a reduced AIC of 176 and two statistically significant estimators at a significance level (alpha) of 0.05

To compare the predictive powers of the first and second models, a likelihood ratio test was conducted to determine which of the two models was more significant.

Likelihood Ratio Test (first model vs. second model)  
A close-up of a math

Description automatically generated

As expected, the test confirmed that first model was not statistically better than the second model in terms of predicting Parkinson’s outcome. With a chi-square value of 0.949 and a relatively negligible change in deviance, this finding further reinforces that income and education did not significantly enhance or contribute to the model's performance. Hence, there is evidence to suggest that, numerically, income and education might not be essential components of the model and were therefore excluded.

Given that the second model with the variables gender and age demonstrated greater statistical significance, a third model was formulated to investigate potential interactions between these two variables.

The interaction term in this model is significant at an alpha level of 0.1. The overall model seems to be statistically significant, indicated by a reduced AIC, diminished residual deviance, and the presence of two statistically significant variables at an alpha significance level of 0.001.

A screenshot of a computer

Description automatically generated

A likelihood ratio test was conducted to compare the predictability of the second and third models – that is, the model without the interaction term and the model with the interaction term.

Likelihood Ratio Test (second model vs. third model)  
A white background with black text

Description automatically generated

With a chi-square value of 0.0621, the test results reveal that the third model, incorporating the interaction term, possesses a higher predictive power than the second model. Furthermore, the drop in residual deviance (166.6) demonstrates that the inclusion of the interaction term improves the model, which supported its inclusion in the final model.

Therefore, the third model, with the interaction term between gender and age, was chosen as the final model.

Now, the overall significance of the chosen final model can be assessed.

A screenshot of a computer

Description automatically generated

The likelihood ratio test reveals a chi-square value of 0.0004, clearly indicating the statistical significance of this model. Furthermore, the c-index, representing the area under the ROC curve, indicates that the model possesses approximately 70% discriminatory power to differentiate between the two categories (in this context, individuals likely to have Parkinson's disease and those unlikely to have Parkinson's disease).

Therefore, the final model is given by:

Utilizing this model will allow us to address the research question: What effect does gender have on the prevalence of Parkinson’s disease? This model suggests that the effect of gender on PD prevalence does indeed vary depending on age.

1. **Process**

The initial stage in the process of building the model was to identify and justify the use of logistic regression in the context of determinants and Parkinson's disease. He, L., et al. (2016), demonstrated the appropriateness of utilizing this modeling approach when they successfully used it to examine the “Key Determinants of Quality of Life in Parkinson's Disease Patients”.

The next step involved the selection of predictor variables that align with the research question and potentially correlate with the prevalence of PD. Gender (GEN), Age (AGE), Education (EDU), and Income (INCOME) emerged as potential determinants based on prior research literature that underscores their potential impact on health outcomes. For instance, studies have consistently explored the likelihood of older age being a significant risk factor for PD (Ball, N., et al., 2019). Moreover, gender differences have also been a subject of study, with men exhibiting higher susceptibility (Miller, I. N., & Cronin-Golomb, A., 2010). Furthermore, factors like education levels and income have shown associations with health disparities, including the prevalence of Parkinson's disease (Najafi, F., et al., 2023)).

Furthermore, the decision to include the interaction term GEN\*AGE is supported in research suggesting that age can modify gender-based health disparities (Schober, P., & Vetter, T. R. (2021)). Ultimately, the final model was chosen after undergoing iterative refinements, which involved considering goodness-of-fit measures such as the likelihood ratio test and the AIC and conducting residual analysis.

1. **Tools**

The model was constructed using the R programming language and its associated statistical software. This tool is appropriate because R provides specific packages and functions designed for modeling logistic regressions, allowing for efficient handling of large datasets, conducting precise calculations, and performing necessary statistical analyses. In addition, many healthcare researchers widely use R for its commendable visualization and communication of results (Jalal, H., *et al.,* (2017)). For this reason, it was selected as an ideal tool for my research and model type as well.

Furthermore, in the process of model building, the software facilitated the comparison of the different models using techniques like the likelihood ratio test and the AIC measures, where the lower the AIC value indicates a better fit of the model.

1. **Analysis**

The final model to predict the prevalence of Parkinson's disease is:

Analysis on Gender:

The model results indicated that, on average, being female (GEN = 1) was associated with an increase of 2.15 in the log-odds of Parkinson's disease prevalence compared to being male (GEN = 0), while holding age constant. This is clearly depicted in the graph below, where the bar for females is notably taller than the bar for males. This suggests that, based on my model, the incidence of PD is higher for females compared to males.

Furthermore, the vertical distance between the two bars reflects the magnitude of the difference in PD incidence between males and females. In this case, the difference is quite significant, implying that gender plays a substantial role in PD prevalence according to my model.

A graph with a blue and pink rectangle

Description automatically generated

Analysis on Age

The model demonstrates that, on average, individuals aged 65 or older (AGE = 1) experience a 0.57 increase in the log-odds of developing Parkinson's disease compared to those under 65 (AGE = 0), while accounting for other factors. Examining the graph below, it becomes evident that the "65 & above" bar is higher than the "below 65" bar. This observation suggests that the model predicts a greater likelihood of Parkinson's disease in the older age group.

A graph showing the age group

Description automatically generated

65 & above below 65

Analysis on the interaction term:

The coefficient of the interaction term (-1.432) indicates the extent of change in the effect of gender (GEN) on the log-odds of Parkinson's disease prevalence for every unit change in age (AGE). Specifically, it signifies how the effect of being female (GEN = 1) on PD prevalence shifts with a one-unit increase in age. This suggests that while being female tends to increase the odds of PD prevalence, this effect decreases as individuals get older. This could imply that gender might have a more significant impact on PD prevalence in younger age groups and a diminishing impact in older age groups.

The graph below illustrates the relationship between the incidence of Parkinson's disease (PD) and age, as captured by the interaction term (GEN\*AGE).

A graph with a line and a red dot

Description automatically generated

The diverging lines suggest that the effect of gender on the odds ratio changes with age. Specifically, being female might have a larger impact on the odds ratio for individuals below 65 compared to those 65 and older.

For individuals below 65 years old (blue points), the odds of PD are higher for females (GEN = 1) compared to males (GEN = 0). For individuals aged 65 and older (red points), the odds of PD are also higher for females (GEN = 1) compared to males (GEN = 0).

In both age groups, the odds ratios for females are greater than 1, indicating a higher odds of PD compared to males.

Answering My research question:

To address the question of how gender impacts the prevalence of Parkinson's disease, I needed to determine the odds ratio for individuals aged 65 and older as well as those below 65. The odds ratios were calculated by exponentiating the coefficients.

For individuals aged 65 and older (AGE = 1), the odds ratio was 2.046232. This indicates that among individuals aged 65 and older, females (GEN = 1) have an approximately 2.05 times higher odds of having Parkinson's disease compared to males (GEN = 0) in the same age group while holding all other variables constant.

For individuals aged below 65 (AGE = 0), the odds ratio was 8.571134. This indicates that among individuals below 65 years old, females (GEN = 1) have 8.57 times higher odds of developing Parkinson's disease compared to males (GEN = 0) in the same age group while holding all other variables constant.

For both age groups, the odds ratio for Gender (GEN) shows that females have a significantly higher odds of having Parkinson's disease compared to males. However, the odds ratios further confirm that gender might have a more significant impact on PD prevalence in younger age groups and a diminishing impact in older age groups

1. **Limitations**

Since the dataset is solely from the greater Boston metropolitan area, it may not fully represent the diverse nature of Parkinson's disease prevalence across different geographical regions. Thus, generalizing my findings to other populations pose limitations. Additionally, the small sample size of 139 observations might limit the model's robustness and affect the precision of coefficient estimates and statistical significance. For this reason, confidence intervals were not considered in my analysis. While confidence intervals are valuable tools for estimating the range of values for odds ratios, their applicability becomes limited and less reliable when dealing with a small sample size.

Additionally, the potential bias in my results could arise from the unequal distribution of individuals with PD and those without PD within different subgroups of my dataset. For example, as seen in the tables below, the total number of individuals with PD (86) compared to those without PD (53), represents an imbalanced distribution, which might lead to a disproportionate influence on the model's predictions. Similarly, considering age groups, those aged above 65 with PD (52) and those without PD (25), could introduce bias by potentially giving more weight to the group with a higher representation. For gender, males with PD (54) greatly outnumber those without PD (16), while the numbers are more balanced for females with PD (32) and those without PD (37). This gender imbalance within the PD group might skew the results.

A screenshot of a computer code

Description automatically generated

Other limitations include the model's inclusion of interaction terms, as this could introduce complexity. While it enhances our understanding, it might not capture every interaction that shapes PD prevalence. Finally, while several determinants were accounted for, other unmeasured variables like environmental pollutants might influence Parkinson's disease prevalence.

1. **Approach**

The first step I took to approach my research question was framing a research question that explored the relationship between social determinants and Parkinson's disease prevalence. As I delved deeper into my research, I decided to specifically focus on the effect of gender and its potential interaction with age on PD prevalence. I then selected a relevant dataset from the ICPSR’s website that contained variables related to Parkinson's disease and potential determinants such as gender, age, education, and income. I prepared the dataset by transforming and recoding variables to ensure their compatibility with the model and research question.

Next, I decided to use a logistic regression model because of the categorical nature of my outcome variable (Parkinson's disease prevalence). This choice was defended by citing literature that successfully utilized and supported the applicability of logistic regression in studying disease outcomes and determinants.

I then started the process of building my model by including gender, age, education, and income as predictor variables in my initial model. I iteratively refined the model using the AIC and likelihood ratio tests, which guided me to exclude education and income due to their lack of statistical significance. This resulted in a simplified model with gender and age as predictors. I then introduced an interaction term (GEN\*AGE) to explore the potential interaction between gender and age.

The likelihood ratio test and the c-index were used to confirm the final model's statistical significance and discriminatory ability, respectively, which supported the validity of the final chosen model with the interaction term.

I analyzed the coefficients of the model to interpret the effects of gender, age, and the interaction term on Parkinson's disease prevalence. I calculated and interpreted odds ratios to quantify the impact of age and gender differences, focusing on the interaction effect. I then incorporated graphs to show the incidences and the interaction term visually.

I acknowledged the limitations inherent in the dataset, my model, and analysis, including potential biases resulting from the dataset's geographic specificity, the sample size, and the exclusion of certain variables.

And finally, I used my finalized model to answer the research question: "What effect does gender have on the prevalence of Parkinson’s disease?" By interpreting odds ratios and considering the interaction term, I concluded that the effect of gender on PD prevalence varies with age.

1. **Applicability**

The purpose of my research was to explore the explore the influence of social determinants on the prevalence of Parkinson's disease. My model was specifically designed to address the research question of how gender impacts Parkinson's disease prevalence and whether this effect varies with age. The model aligns well with the research question by quantifying and providing insights into the effects of gender and age on Parkinson's disease incidence.

The model helps answer the research question by offering estimates of how being female impacts the odds of having Parkinson's disease compared to being male, while considering the age factor. The interaction term enabled the assessment of whether this gender effect changes across different age groups.

The model's findings suggests that gender is indeed associated with the odds of Parkinson's disease, with females having a significantly higher odds ratio compared to males. Additionally, the interaction term shows that the effect of gender on Parkinson's disease prevalence varies depending on age. Furthermore, even though existing research has found a higher risk of males affected by PD, my findings of gender reflect some previous research that suggests a higher prevalence of Parkinson's disease in females (Nina Avramova, 2018). Additionally, the findings for Age are also consistent with the very known association between increasing age and Parkinson's disease risk (Collier, T. J., Kanaan, N. M., & Kordower, J. H. (2017)).

References

Ball, N., Teo, W. P., Chandra, S., & Chapman, J. (2019). Parkinson's Disease and the Environment. *Frontiers in neurology*, *10*, 218. <https://doi.org/10.3389/fneur.2019.00218>

Collier, T. J., Kanaan, N. M., & Kordower, J. H. (2017). Aging and Parkinson's disease: Different sides of the same coin?. *Movement disorders : official journal of the Movement Disorder Society*, *32*(7), 983–990. <https://doi.org/10.1002/mds.27037>

Emergence and Evolution of Social Self-management of Parkinson's Disease, Greater Boston Metropolitan Area, 5 states, 2013-2019. *ICPSR.* [*https://archive.icpsr.umich.edu/view/studies/37631/explore-data*](https://archive.icpsr.umich.edu/view/studies/37631/explore-data)

He, L., Lee, E. Y., Sterling, N. W., Kong, L., Lewis, M. M., Du, G., Eslinger, P. J., & Huang, X. (2016). The Key Determinants to Quality of Life in Parkinson's Disease Patients: Results from the Parkinson's Disease Biomarker Program (PDBP). *Journal of Parkinson's disease*, *6*(3), 523–532. <https://doi.org/10.3233/JPD-160851>

 Hoseinipalangi, Z., Kan, F. P., Hosseinifard, H., Doustmehraban, M., Masoumi, M., Rafiei, S., Barmayoon, P., Ahmadi, N., Dehnad, A., Eshtod, H., Asl, M. T., Hoseini, B. S., Nasiri, M., Arian, M., Aghalou, S., & Ghashghaee, A. (2023). Systematic review and meta-analysis of the quality-of-life of patients with Parkinson’s disease. *Eastern Mediterranean Health Journal*, *29*(1), 63–70. <https://doi-org.ezproxy.snhu.edu/10.26719/emhj.23.013>

Jalal, H., Pechlivanoglou, P., Krijkamp, E., Alarid-Escudero, F., Enns, E., & Hunink, M. G. M. (2017). An Overview of R in Health Decision Sciences. *Medical decision making : an international journal of the Society for Medical Decision Making*, *37*(7), 735–746. <https://doi.org/10.1177/0272989X16686559>

Mesa J. L. (2004). Understanding data in clinical research: a simple graphical display for plotting data (up to four independent variables) after binary logistic regression analysis. *Medical hypotheses*, *62*(2), 228–232. <https://doi.org/10.1016/S0306-9877(03)00335-9>

Miller, I. N., & Cronin-Golomb, A. (2010). Gender differences in Parkinson's disease: clinical characteristics and cognition. *Movement disorders : official journal of the Movement Disorder Society*, *25*(16), 2695–2703. <https://doi.org/10.1002/mds.23388>

Najafi, F., Mansournia, M. A., Abdollahpour, I., Rohani, M., Vahid, F., & Nedjat, S. (2023). Association between socioeconomic status and Parkinson's disease: findings from a large incident case-control study. *BMJ neurology open*, *5*(1), e000386. <https://doi.org/10.1136/bmjno-2022-000386>

Nina Avramova (October 2, 2018). Nearly half of women and a third of men will develop dementia, stroke or Parkinson’s, study says. CNN <https://www.cnn.com/2018/10/02/health/dementia-parkinsons-stroke-risk-study-intl/index.html>

Park, R. S., Rattana-Arpa, S., Peyton, J. M., Huang, J., Kordun, A., Cravero, J. P., Zurakowski, D., & Kovatsis, P. G. (2021). Risk of Hypoxemia by Induction Technique Among Infants and Neonates Undergoing Pyloromyotomy. *Anesthesia and analgesia*, *132*(2), 367–373. <https://doi.org/10.1213/ANE.0000000000004344>

Schober, P., & Vetter, T. R. (2021). Logistic Regression in Medical Research. *Anesthesia and analgesia*, *132*(2), 365–366. <https://doi.org/10.1213/ANE.0000000000005247>

Wan, Neng & Kan, Ge. (2015). Parkinson's Disease and Pesticides Exposure: New Findings From a Comprehensive Study in Nebraska, USA. *The Journal of Rural Health*. 32. n/a-n/a. 10.1111/jrh.12154. <https://www.researchgate.net/publication/283974651_Parkinson's_Disease_and_Pesticides_Exposure_New_Findings_From_a_Comprehensive_Study_in_Nebraska_USA>

Wirdefeldt, K., Adami, H. O., Cole, P., Trichopoulos, D., & Mandel, J. (2011). Epidemiology and etiology of Parkinson's disease: a review of the evidence. *European journal of epidemiology*, *26 Suppl 1*, S1–S58. <https://doi.org/10.1007/s10654-011-9581-6>

Yang, F., Johansson, A. L. V., Pedersen, N. L., Fang, F., Gatz, M., & Wirdefeldt, K. (2016). Socioeconomic status in relation to Parkinson's disease risk and mortality: A population-based prospective study. *Medicine*, *95*(30), e4337. <https://doi.org/10.1097/MD.0000000000004337>

Yang, W., Hamilton, J.L., Kopil, C. *et al.* (2020). Current and projected future economic burden of Parkinson’s disease in the U.S.. *npj Parkinsons Dis.* **6**, 15. <https://doi.org/10.1038/s41531-020-0117-1>