Fox Insight database and machine learning to study differentiating factors by sex in Parkinson Disease

Analysis of differences in women with Parkinson’s who have taken hormonal replacements

Maitane Gomez Gonzalez  
UPV/EHU  
San Sebastian Basque Country Spain  
mgomez168@ikasle.ehu.eus

Olatz Arbelaitz Gallego  
Computer Architecture and Technology  
UPV/EHU  
San Sebastian Basque Country Spain  
olatz.arbelaitz@ehu.eus

ABSTRACT
Parkinson’s disease (PD) is currently the second most common neurodegenerative disorder and manifests in both motor and non-motor symptoms. Its complexity lies in its diverse manifestations and the variety of symptoms it can present [1].

The literature demonstrates differences in sex due to physiopathological, environmental, and genetic factors, which may influence susceptibility to various diseases, including neurodegenerative disorders such as Parkinson’s. Unfortunately, these variations have largely been overlooked in preclinical and clinical research stages, possibly due to resistance to using female animals in preclinical studies and the low participation of women in clinical trials, where their access is limited. A clear example of this lack of consideration is evident in the report on drug withdrawals conducted in the United States between 1997 and 2000 by the Food and Drug Administration, which revealed that 8 out of 10 withdrawn drugs posed greater risks to women [5]. On the other hand, explainable machine learning techniques can contribute to discover sex and gender differences in medical contexts [5][6].

Fox Insight is an online self-report database created by Michael J. Fox Foundation in 2017 to collect information about PD [4]. It emerged to address the lack of access to diverse information in Parkinson’s research, caused by limitations in minority representation. While concerns may arise regarding the reliability of projects utilizing self-report databases, the comparative study "Comparison of an Online-Only Parkinson's Disease Research Cohort to Cohorts Assessed In Person" conducted by Chahine et al. in 2020 [2] concluded that research using self-reported online data on PD is not only feasible but also exhibits characteristics similar to those carried out in person.

1 OBJECTIVE
This project has three purposes: to examine the content and opportunities provided by FOX Insight for PD research, objectively extracting data for two complete databases without missing values, including the maximum possible number of variables and patients; to determine if PD affects men and women differently; and to analyze the potential impact of taking any form of hormone replacement during or after menopause for at least six months on women with Parkinson’s.

2 ANALYSIS OF FOX INSIGHT DATABASE
The Fox Insight database, available at https://foxden.michaeljfox.org/, houses 125,000 registered patients and features a battery of over 7,000 questions. However, the complexity in its performance and management complicates any research attempts.

The database’s current design has limitations in data access and manipulation due to inefficient filtering. Manual navigation through 7000+ variables is necessary, with redundancy in questions and complications in handling multiple-choice responses. This, alongside a lack of data normalization, complicates preprocessing. The extensive structure of the questionnaire limits user overlap, making it difficult to generate databases and highlighting once again that in data mining processes, 80% or more of the time is invested in the data preparation and preprocessing stages.

3 CREATION OF DBS AND METHODOLOGY
Two databases have been generated with PD patients. One is a general db that includes patients of both sexes with the maximum number of common questions answered. The other is specific to women, containing additional information. Both were created using the same algorithm, the algorithm initially selects four variables (Age, Sex, Education and InitPDDiag, the initial PD diagnostic positive for everyone in this study) and removes null values and duplicates. Then, sequentially incorporates variables based on their non-null size, ensuring a minimum size of 1000, and reverses steps if adding a variable reduces the size below that threshold, with the only difference being that in the specific db, the target variable was added, and all male and female patients without a response in it were removed.
In the general db (189 variables, 1526 patients, 799 men, 727 women), sex has been identified as the dependent variable with the aim of analyzing the presence or absence of differentiating factors by sex in PD patients. In the specific db, the variable determining whether a patient has taken hormone replacements during or after menopause for at least six months was chosen as the study objective, involving 163 variables and 128 patients, with 68 having taken them and 60 not.

Classifiers and explainers were then used to discriminate between the study groups and try to find differences among them. Experiments were carried out with four algorithms, including Random Forest, XGBoost, SVM, and linear perceptron, using a 10-fold cross-validation methodology as validation strategy. For each algorithm, two feature selection heuristics were used to reduce the complexity of the classifiers - Random Forest Feature Selection and CFS using the SelectKBest algorithm. In each classifier, a study of the heuristics was carried out to determine the optimal threshold in RF and the ideal number of variables to choose in SelectKBest. SHAP (Shapley Additive Explanations) [3] based on Shapley Values was then used to explain the results of the classifiers.

4 RESULT OF THE ANALYSES

In the general study, SVM achieved the highest accuracy. When discriminating sex it achieved a 75% combined with the RFSS heuristic. For the specific study db, the linear perceptron with correlation heuristic achieved a 77% rate. In both cases, the variables that have the greatest impact on the classifier’s decision and how they affect the different classes have been analyzed.

The SHAP values for the general db where class 0 represents women, and class 1 represents men can be seen in Figure 1 graph (a).

\[\text{Figure 1: SHAP results for (a) general study (b) specific study}\]

According to the classifier, it tends to favor women in situations where anxiety (NonMoveAnxious), self-esteem (MoodWorth), muscle pain (LivePDPain), and swallowing or chewing complications (NonMoveSwall) are higher. On the other hand, it leans towards men when a higher level of urinary incontinence (NonMoveUrinePM) and a lesser change in sexual desire (NonMoveSex) are recorded.

The results of the specific db where, class 0 represents having taken a type of hormonal replacement, and class 1 represents not having taken them is shown in Figure 1 graph (b).

The classifier considers that the patient has taken hormonal replacement if they have higher urinary incontinence (NonMoveUrine), lesser change in sexual desire, fewer difficulties in staying awake (NonMoveAwake), and a greater feeling of depression (NonMoveFeel).

5 CONCLUSIONS

The Fox insight db is hard to manage in its current form and limits the number of people filling it. Simplifying it certainly would help. However two databases could be extracted and in both cases some differences were suggested. These findings emphasize the importance of considering sex differences and hormone therapies in managing Parkinson’s symptoms, shaping future research for personalized treatment approaches. Furthermore, computational techniques play a crucial role in unbiased research, facilitating efficient analysis of large datasets.

ACKNOWLEDGMENTS

This work was funded by Grant PID2021-123087OB-I00 funded by MCIN/AEI/10.13039/501100011033 and, ERDF a way of making Europe and the Department of Economic Development and Competitiveness (ADIAN, IT1437-22) of the Basque Government.

The Fox Insight Study (Fi) is funded by The Michael J. Fox Foundation for Parkinson’s Research. We would like to thank the Parkinson’s community for participating in this study to make this research possible.

REFERENCES


