Group 8 Proposal

Jingnan Bai, Kexin Wang, Yiming Miao, Qiyu Huang

2024-02-21

SPECFIC AIMS

Parkinson's disease (PD) ranks as the second-most common neurodegenerative disorder in the US, affecting nearly 90,000 individuals annually. Characterized by uncontrollable movements such as shaking, stiffness, and imbalance, PD significantly diminishes patients' quality of life.

Characterized as a progressive neurological disorder, PD occurs due to the gradual degeneration or death of nerve cells in the basal ganglia. The exact cause of neuronal death remains inconclusive, though current research suggests a combination of specific genetic variants and environmental factors, including toxin exposure, as leading risk contributors.

The UK Biobank is a large biomedical database that contains detailed genetic, clinical, and lifestyle information on 500,000 UK participants. Notably, it includes precise records of PD diagnoses among participants, offering a unique lens to longitudinally examine the characteristics of PD patients and hence deepen our understanding of the disease's onset.

This research project, informed by a thorough review of the dataset and existing medical literature, is driven by two specific aims:

A. Specific Aim 1

To analyze the influence of patient information on PD occurrence. Leveraging the UK Biobank's extensive data on demographics, habits, diets, and environmental exposures, we aim to construct a statistical model that explains the relationship between these variables and PD occurrence, identifying factors with significant inference.

B. Specific Aim 2

To develop predictive models for early PD onset. Beyond demographic and lifestyle data, the UK Biobank provides extensive information on participants' limbs and trunks. Utilizing all these data, our objective is to create a predictive model with high accuracy to help the early detection of PD.

We anticipate that this project will uncover factors that elucidate or predict the onset of PD. We hope such findings could enable individuals at higher risk to benefit from earlier interventions and contribute to the well-being among aging people.

RESEARCH STRATEGY

A. Significance

Analyzing Parkinson's disease (PD) data is essential to improving our understanding of this complex neurodegenerative disease and improving both diagnosis and treatment. Through the examination of PD-related information, researchers can identify not-so-obvious patterns, including genetic, environmental, and lifestyle factors that influence the disease's development. This insight is instrumental in creating models to detect individuals at increased risk sooner, facilitating early actions that can profoundly modify the course of PD. Additionally, by identifying biomarkers for the early detection and monitoring of Parkinson's disease (PD), treatment plans can be more specific. It also advances the development of novel treatments by illuminating the basic mechanisms underlying Parkinson's disease. Ultimately, analyzing PD data holds promise for improving the quality of life for those impacted worldwide and for sparking hope for improved treatment and a cure.

B. Innovation

We face a significant challenge in our research on Parkinson's disease (PD) because of the naturally skewed dataset, which has much fewer PD patients than the healthy population. This imbalance might skew our analysis and produce less trustworthy findings. Our study uses methods intended to efficiently balance the dataset in order to address this problem in a novel way. Techniques like undersampling the control group, oversampling PD cases, or using synthetic data generation, such as SMOTE (Synthetic Minority Over-sampling Technique), are used. By normalizing the dataset distribution, these techniques help to improve the accuracy of our analysis and make it more representative of the true impact of PD. Our goal is to improve the basis for future PD research by proactively addressing this imbalance and gaining insights that are both applicable and valid.

In terms of model selection, our methodology goes beyond conventional techniques by incorporating a range of predictive models, such as Lasso Regression, Ridge Regression, and Logistic Regression, each of which is known for its special abilities in managing particular kinds of data features. To bolster the robustness and accuracy of our predictions, we also integrate advanced ensemble techniques like BOOSTING and BAGGING. These methods, by pooling predictions from multiple models, significantly enhance our analytical framework's performance. They help in reducing the risk of overfitting and increasing model reliability. This comprehensive and innovative modeling strategy not only elevates the precision of our PD predictions but also deepens our understanding of the disease's intricate patterns, offering a solid basis for developing more effective diagnostic and therapeutic interventions.

C. Research Plan

We present our research on Parkinson's disease (PD) in four condensed steps for maximum efficiency and clarity. The first step of the process is data processing, where we deal with the imbalance in our dataset, which is a typical problem in medical research. We guarantee fair representation between PD cases and controls by using methods like SMOTE for oversampling and strategic undersampling, providing a foundation free of bias for our analysis.

Transitioning to the exploratory data analysis (EDA) phase, our goal is to sift through the data for preliminary insights and identify potential PD predictors. This critical step not only helps us grasp the dataset's underlying structure but also steers the direction of our deeper investigation.

We use complex algorithms in the feature importance analysis that follows to identify the variables that have the biggest impact on the risk of Parkinson's disease. This stage is critical because it reveals the relative importance of each feature and their causal relationships, deepening our comprehension of the complex nature of PD. The understanding we gain from this experience will be crucial to interpreting our results in a meaningful and persuasive way.

Our work culminates in the model building stage, where we use multiple predictive techniques, such as Ridge, Lasso, and Logistic Regression, as well as ensemble approaches like BOOSTING and BAGGING. Our models are put through a rigorous evaluation process to make sure they have strong generalization capabilities in addition to being an accurate predictor of PD risk. This rigorous and concise methodology ensures a targeted investigation of PD data, with the goal of providing deep insights into its forecast and broad management approaches.

D. Specific Aim 1: To investigate the combined effects of patient demographics, dietary patterns, and lifestyle habits on the occurrence of Parkinson's disease.

D.1. Hypothesis

Patient demographics (age, gender, ethnicity), diet (specific dietary patterns or nutritional intake), and habits (exercise frequency, tobacco use, alcohol intake) can significantly affect the occurrence of Parkinson's disease.

D.2. Rationale

The emphasis on these factors is due to the multifactorial nature of Parkinson's disease, which involves complex interactions between genetic predisposition and environmental exposures. In previous clinical research, it has been established that specific dietary components such as antioxidants found in fruits and vegetables can reduce oxidative stress, which is crucial in the neurodegenerative process observed in Parkinson's disease. While these results provide valuable information on protective dietary elements, a key question that remains in our understanding is how broader lifestyle patterns, including

exercise, tobacco use, and alcohol consumption, interact with dietary habits and demographic backgrounds to influence the risk of Parkinson's disease.

D.3. Experimental Approach

To explore potential significant features, we will conduct data processing, data visualization, and statistical analysis based on the following steps:

- Data Preparation: Missing data will be handled based on the nature and extent of the missingness. Dietary intake and exercise data will be converted into standardized units such as servings per day, and hours per week to enable consistent comparisons. Additionally, we may also normalize variables with skewed distributions to satisfy the assumptions of certain statistical tests.
- Descriptive Statistics: Summary statistics (mean, median, standard deviation, frequencies) will be calculated to provide a comprehensive overview of the data and identify potential issues for further investigation.
- Data Visualization: We will employ individual plots or heatmaps for interaction effects to visualize the relationship between factors and the outcome.
- Statistical Model: Recognizing the challenge of multicollinearity, where closely related factors may interface with the ability to discern their individual effects, we will first select potential factors in each category. We may consider using models that include penalties such as LASSO or ridge regression. Subsequent to feature selection, we will conduct statistical tests to assess the significance of these factors. As longitudinal measurements for each feature are available, we will implement mixed-effect models to capture the within-participant variability over time and provide an understanding of how individual trajectories influence the outcome.
- Sensitivity Analysis: We will perform sensitivity analyses to validate model assumptions and ensure our findings regarding key factors are reliable.

D.4. Interpretation of Results

Given the results from the statistical models, we will interpret the results in the context of existing literature, and how our results align with or diverge from previous studies. We will also discuss the practical implications for the prevention and management of Parkinson's disease.

D.5. Potential Problems and Alternative Approaches

Unaddressed confounding variables, which influence both the exposure (diet, lifestyle habits) and the outcome (Parkinson's disease), may bias the results. Socioeconomic status, for example, may impact participants' access to healthcare resources and safe environments for exercise, which could affect the occurrence of the disease. A potential solution involves adjusting for other known confounders in the multivariate model.

E. Specific Aim 2: Establishing practical and interpretable models for predicting the risk of Parkinson's disease.

E.1. Hypothesis

Parkinson's disease is widely believed to be associated with a combination of various risk factors, making the prediction of the disease's risk feasible. Based on previous detection and analysis, we will utilize interpretable predictive models to further validate identified Parkinson's-related features, while exploring more about indicators that may not be causative but could assist in early Parkinson's diagnosis.

E.2. Rationale

While no definitive cause has been identified in previous studies, various features have been linked to the risk of Parkinson's disease, including genetics, environmental risk factors, lifestyle, and family history of related syndromes. Given the potential interaction among these factors, early diagnosis of Parkinson's disease could be complex and challenging, which denotes that more comprehensive methods are required to help model complex scenarios, capture relationships, and achieve more accurate risk predictions.

E.3. Experimental Approach

To construct a practical risk prediction model, we will conduct data processing, model training, and evaluation based on the following steps:

- Data Preparation: Involving data cleaning, checking for missing data, and basic preprocessing to ensure data quality. Considering the large number of features in the dataset, we will utilize data compression and other efficient feature filtering methods to reduce model complexity.
- Feature Engineering: Exploring the relationship between features and the response variable with exploratory data analysis and basic visualization tools, and conducting appropriate feature engineering such as binning, encoding and spline tricks if needed. Additionally, considering the data imbalance, we will focus on sampling techniques and related evaluation in data processing to assist the model learning process.
- Model Training and Improvement: Utilizing various ensemble learning methods (boosting or bagging) to construct predictive models for estimating the risk of Parkinson's disease. We will further improve model performance with appropriate parameter tuning and regularization strategies, considering the potential problem of overfitting with imbalanced data and sampling preprocessing.
- Model Evaluation: With cross-validation techniques, further assessing the predictive performance of the models based on basic metrics for classification tasks, along with AUC for evaluating the model's ranking abilities. Furthermore, given the data characteristics, more attention will be attached to examining the robustness of the models when facing the imbalanced data in practical use.
- Interpretation: Conduct model selection comprehensively based on multiple metrics. Calculating feature importance using ensemble learning methods to complement and support the conclusion of causal analysis in the previous section. Also, analyze indicators that may not have causality but could assist in Parkinson's disease diagnosis, providing insights into its early detection.

E.4. Interpretation of Results

The predictive model will focus on a balance between accuracy and interpretability, which would provide a new perspective on examining the risk factors of Parkinson's disease. With the analysis of feature importance and the information value of each feature, we will further identify features that have causal relationships with Parkinson's disease risk or could assist in early detection of Parkinson's disease.

E.5. Potential Problems and Alternative Approaches

Given the extreme class imbalance and the complex interactions among risk factors of Parkinson's disease, oversampling techniques like SMOTE may introduce ambiguity to classification boundaries and exacerbate potential confusion and overfitting. To address the possible challenges, we plan to use cost-sensitive learning algorithms and other methodologies as alternative approaches to mitigate these issues.

REFERENCE

Parkinson's Disease: Challenges, Progress, and Promise. National Institute of Neurological Disorders and Stroke. Updated Jan 30, 2023. Accessed Feb 21, 2024. https://www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise

Parkinson's Disease. Mayo Clinic. Updated May 26, 2023. Accessed Feb 21, 2024. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055

Parkinson's Disease: Causes, Symptoms, and Treatments. National Institute on Aging. Updated April 14, 2022. Accessed Feb 21, 2024. https://www.nia.nih.gov/health/parkinsons-disease/parkinsons-disease-causes-symptoms-and-treatments

What is UK Biobank?. UK Biobank. Accessed Feb 21, 2024. https://www.ukbiobank.ac.uk/

View the Data Showcase. UK Biobank. Accessed Feb 21, 2024. https://biobank.ctsu.ox.ac.uk/crystal/

Park HA, Ellis AC. Dietary Antioxidants and Parkinson's Disease. Antioxidants (Basel). 2020;9(7):570. Published 2020 Jul 1. doi:10.3390/antiox9070570

Chairta PP, Hadjisavvas A, Georgiou AN, et al. Prediction of Parkinson's Disease Risk Based on Genetic Profile and Established Risk Factors. Genes (Basel). 2021;12(8):1278. Published 2021 Aug 20. doi:10.3390/genes12081278

Yoon SY, Park YH, Lee HJ, Kang DR, Kim YW. Lifestyle Factors and Parkinson Disease Risk: Korean Nationwide Cohort Study With Repeated Health Screening Data. Neurology. 2022;98(6):e641-e652. doi:10.1212/WNL.0000000000012942