The Thesis Committee for Jingpeng Zhai
certifies that this is the approved version of the following thesis:

Causal inference for investigating Parkinson's disease pathogenesis

SUPERVISING COMMITTEE:

Chandrajit Bajaj, Supervisor

Alexander Huth

John Virostko

Causal inference for investigating Parkinson's disease pathogenesis

by

Jingpeng Zhai

Thesis

Presented to the Faculty of the Graduate School of The
University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Computer Science

The University of Texas at Austin
December 2023

Epigraph

"...observational studies are an interesting and challenging field which demands a good deal of humility, since we can claim only to be groping toward the truth."

— Cochran, 1972

Acknowledgments

This is to everyone who gave me a chance. Special thanks to my supervisor Dr. Bajaj for introducing me to the field of causal inference and for providing guidance.

Abstract

Causal inference for investigating Parkinson's disease pathogenesis

Jingpeng Zhai, MSCS
The University of Texas at Austin, 2023

SUPERVISOR: Chandrajit Bajaj

Randomized control trials have long been regarded as the standard method for establishing causal relationships. However, in situations where it is impractical to carry out such trials, observational studies involving natural & random variations along with causal inference methods can be used to reason about causality. Causal inference methods require the expression of expert domain knowledge in the form of a causal model. But what happens in situations where there is little to no prior knowledge? In a dataset with a plethora of variables, how should one identify & isolate potential treatment and outcome variables? For example, Parkinson's disease (PD) is a disorder with diverse manifestations, multiple proposed molecular pathways but no established etiology. Given a dataset with PD patients and healthy controls, and their clinical data ranging from varying levels of biology, how does one approach causal graph construction? In this paper, we devise a scheme that uses gradient-boost tree ensemble algorithms to identify systematically important features for use in causal graph construction, and attempt to establish causal relationships between them based on biological hierarchy. Lastly, we find one genotype feature of α-synuclein to have a significant causal effect on PD diagnosis.

Table of Contents

Chapter 1: Introduction

It is said that "most applied science is concerned with uncovering causal relationships" (1). The establishment and measurement of causality are usually done through randomized control trials (RCTs). RCTs are a type of scientific experiment that attempts to evaluate treatment effects by assigning participants a treatment by a chance mechanism while controlling for all other variables (2). RCTs have been utilized in a wide range of fields including psychology, education, social sciences, public policy, and economics (3; 4; 5; 6; 7). Well-defined and rigorously executed RCTs with a large group of participants can provide useful information regarding the treatments of interest and remain the most robust method for quantifying causal relationships (8). An example of RCT is the determination of the efficacy and safety of the BNT162b2 mRNA vaccine against SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) and its resulting disease Covid-19. The experimentalists designed a multinational, placebo-controlled, randomized, observer-blinded immunogenicity & safety trial during which 43,548 participants were randomly given either the treatment (BNT162b2 mRNA vaccine) or a saline placebo (trial NCT04368728 at `ClinicalTrials.gov`). Then, the participants were monitored for a median of 2 months with their Covid-19 status and general health status data recorded (9). Afterward, when analyzing the data between the treatment and placebo group, scientists found a two-dose regimen of the BNT162b2 vaccine is 95% effective in the prevention of Covid-19 with a 95% confidence interval of $[90.3, 97.6]$. Further, when broken down into subgroups by age, sex, race and pre-exisitng conditions, the vaccine still had similar efficacies of 90%+ (9). This RCT and its data provided concrete evidence demonstrating that immunization can prevent Covid-19 and validated the RNA-based vaccine technology. It informed public health policies that resulted in tens of millions of lives saved in the 1st year of vaccination alone (10).

However, even with its advantages including a long history of development,

well-characterized protocols, and an immensely positive social impact, RCTs are not without their limitations and drawbacks. Some disadvantages include the high cost of implementing RCTs, the long time horizon needed to evaluate certain interventions, ethical issues regarding giving certain participants adverse treatments, only 1 or a few variables can be studied at the same time in an RCT, making it unlikely to capture the full picture of complicated phenomena and lastly, the practical constraints for many questions to be answered through RCTs (11; 12). For example, what is the effect of school closures on student learning? How do minimum wage laws affect the employment rate? What is the effect of immigration on wages? (13; 1) To answer these types of questions, and address some of the shortcomings of RCTs, researchers employ "observational studies" to uncover, quantify, and verify the cause-and-effect relationships. Observational studies or "natural experiments" take advantage of the natural & random variation in data to reason about causality – such is the field of causal inference.

For this study, by using the example of Parkinson's disease (PD) – a disease with unknown etiology and diverse manifestations, we aim to devise strategies for identifying potential causal variables and then assessing & validating their treatment effect on disease biomarkers to uncover insights regarding PD pathogenesis.

Chapter 2: Parkinson's disease

2.1 Parkinson's disease overview

Parkinson's disease or PD is an age-related neurodegenerative disease that poses a serious health and socioeconomic challenge for society. Over 6 million patients worldwide have been affected since 2016 with PD being second only to Alzheimer's disease as the most popular idiopathic neurological disorder (14). The disease has long been characterized by its notable motor symptoms which are presumably associated with Lewy bodies and the depletion of dopaminergic neurons in the substantial nigra. However, the newest diagnostic guidelines acknowledge the diverse symptomology of PD and also note the clinical significance of non-motor symptoms. Similarly, the disease's molecular pathology has also evolved beyond Lewy bodies to other protein aggregates, molecular chaperones, neurotransmitters, and prions..etc (15; 16). Studies indicate that PD affects all regions of the world without noticeable epidemiological differences, however, the number of incidences and deaths & disabilities from PD more than doubled over the past two decades (14). Additionally, as the global population ages, PD prevalence is projected to double again over the next 30 years (17; 18). An analysis of data from medical administrative claims, the Medicare Current Beneficiary Survey and Medical Expenditure Panel Survey estimates a US prevalence of 1.04 million individuals with PD in 2017 corresponding to a cumulative economic burden of $51.9 billion. This figure is expected to jump to $79 billion by the year 2037 when there will be an estimated 1.64 million US PD patients (19). However, currently, there are no effective treatments or therapies for PD prevention, to halt PD progression or restore depleted dopaminergic neurons (15).

The projected dramatic increases in PD cases and the associated societal and economic burden of caring underpins the need for the identification of PD etiology and disease mechanisms and subsequent effective treatments and preventative measures to address the said mechanisms.

2.2 Parkinson's Disease pathogenesis

The molecular hallmark pathology of PD includes the accumulation of misfolded amyloid proteins known as Lewy Bodies (LB) in intracellular spaces of dopaminergic neurons of the substantial nigra pars compacta (SNpc), the subsequent dopaminergic neuronal loss in SNpc and exhaustion of dopamine levels in the basal ganglia. LB are protein aggregates comprised of primarily α-synuclein (SNCA), then phosphorylated tau (p-tau), and amyloid-beta (Aβ) protein (14; 15; 16). Other molecular markers of PD include the *parkin* gene, the *DJ-1* dimer, the *PINK1* (PTEN-induced putative kinase), mitochondrial DNA (mtDNA) mutation...etc., their known function and disease mechanism are reviewed in (15).

The clinical manifestations of PD broadly include 2 categories: motor and non-motor symptoms. Motor symptoms can be further divided into 2 categories. Primary motor symptoms include resting tremor, rigidity, bradykinesia and postural instability — the 1st three are widely regarded as the clinical hallmark of PD (20; 15). Secondary motor symptoms include difficulty in using muscles of the mouth, dystonia, sexual dysfunction, and difficulties with voice control. Non-motor symptoms can also be categorized into 2 groups. Primary non-motor symptoms include depression, cognitive dysfunction, muscle & joint pain, sleep insomnia, and general fatigue, while secondary non-motor symptoms include higher incidences of cardiovascular diseases, excessive sweating & oil production from the skin, constipation & urinary problems, and general joint & muscle pain. Additionally, early symptoms of PD can include mild versions of the motor- and non-motor symptoms listed (15).

Despite the well-known molecular and clinical manifestations of PD, the causes of PD remain elusive. Several proposed causes for PD origins include impairment of the *Ubiquitin-Proteasome System* (UPS) pathways, breakdown of *molecular chaperones (heat shock proteins)*, malfunction of the *autophagy lysosomal pathway*. Besides genetic factors, environmental factors such as the toxicity of herbicides and fungicides have also been implicated in the causation of PD (15).

To conclude, PD seems to be a result of complex interactions of both genetic and environmental factors, affecting multiple different molecular pathways, thus resulting in extremely heterogeneous symptoms. Complexities regarding PD etiology must be overcome to facilitate the development of effective & targeted PD treatment.

2.3 Current methods for diagnosing Parkinson's disease

Despite the well-characterized molecular & clinical presentations of PD and diagnostic guidelines from the International Parkinson and Movement Disorder Society, PD can be challenging to diagnose. The PD diagnostic challenges can be attributed to several factors. Firstly, many of PD symptoms, especially at the initial disease stages are atypical and can be related to normal aging. For example, symptoms such as constipation, REM (rapid eye movement) sleep behavior disorder, depression and muscle & joint pain may well be a part of normal aging (15; 14). Secondly, although the detection of LB represents 1 of the most conclusive ways to diagnose PD, this procedure can be only carried out via the examination of post-mortem brain tissues. However, LB has also been found in the brains of individuals with no other signs of PD or Parkinsonism. Thus, if LB were to be utilized as the sole diagnostic indicator, this may result in many false positives (21; 22). Thirdly, PD is remarkably heterogeneous with regard to potential etiologies, symptoms, patterns of progression, comorbidity and treatment response. As a result of this, PD can be confused with other neurodegenerative diseases such as progressive supranuclear palsy and corticobasal degeneration. Several subtypes and spectrums of PD have been also proposed, including genetically defined and sporadic forms of PD (15; 17; 16).

Officially, the UK Parkinson's Disease Society Brain Bank has developed a 3-stage clinical diagnostic criteria for PD. Stage 1 qualification involves the presence of bradykinesia as well as 1 of the following 3 features including muscular rigidity, 4 6 HZ tremor and postural instability. Stage 2 qualification involves exclusion from a list of 16 criteria that may indicate a differential diagnosis. Stage 3 consists of

8 criteria, 3 of which are related to motor symptoms, 2 of which relate to disease progression, and 3 more are related to patient responses to the drug Levodopa. Stage 3 qualification and the diagnosis of PD involves fulfilling at least 3 of the 8 proposed criteria (23). A clinical study with a sample of 100 found the above criteria have a sensitivity of 90%. However, these criteria are generally not used anymore (17).

In recent years, the International Parkinson and Movement Disorder Society (MDS) has devised a new criteria called the MDS-PD criteria. The MDS-PD criteria also employs a 2-step criteria. The first step involves the diagnosis of parkinsonism based on expert neurological assessment of bradykinesia and meeting at least 1 additional cardinal motor symptom. If parkinsonism is confirmed, for the second step, depending on which supportive and exclusion features are met, the patient is further delineated into "clinically established PD" or "clinically probable PD" (24; 17). A validation study on the efficacy of the new MDS-PD criteria found it to have 96% sensitivity and 95% specificity for the identification of "clinically probable Parkinson's disease". While it had a 98.5% specificity and only 59.3% sensitivity for identifying "clinically established Parkinson's disease". Further testing revealed this method had a reduced specificity of 87% for patients with milder manifestations of symptoms (as indicated by <5 years since disease diagnosis) (25). Notably, these criteria use only symptoms and responses to drugs – no explicit molecular or genetic marker is utilized.

Current PD diagnostic tools are inadequate as evidenced by the average one-decade delay between a person's initial symptoms to an official diagnosis and the up to 15% diagnostic error for PD (14). Other than the diagnostic challenges listed above, the newest diagnostic guidelines do not take into account genetic, molecular and imaging tools tools that may contain unique &useful insights not expressed at the tissue or organ level. All of this signals a need for the development of new PD diagnostic approaches, methods for combining useful information from different modalities of tests (e.g., genetic tests, molecular tests, imaging tests, and symptomology), and novel biomarkers that better capture information with regard to disease presence and progression.

2.4 Parkinson's Progression Markers Initiative (PPMI)

The Parkinson's Progression Markers Initiative (PPMI) is an observational, longitudinal study that comprises 423 untreated Parkinson's Disease (PD) patients, 196 healthy controls (HC) and 64 SWEDD (scans without evidence of dopaminergic deficit). The aim of PPMI is to improve the understanding of disease etiology by identifying PD progression markers and establishing biomarker-defined cohorts. It does this by collecting subject data from 4 modalities, including clinical, imaging, genetic and biospecimen data (26). For the purpose of this study, we analyze the PPMI clinical data which in addition to patient information & demographics data, also contains relevant data at varying levels of granularity, from genotype data, to protein data, to motor- and non-motor symptom data. A more detailed exploration of PPMI clinical data is available in section 3.1.

In conclusion, given the projection of doubling of PD burden in the next 3 decades coupled with the very heterogeneous nature of PD, both in its molecular pathways and symptoms, together with an unknown etiology and a recently available large, comprehensive, and longitudinal dataset such as PPMI makes PD a prime candidate for investigation and the application of machine learning & causal inference methods for biomarker discovery and subsequent treatment development.

Chapter 3: Methodology

3.1 Dataset: PPMI clinical dataset

The PPMI clinical dataset which is in heterogeneous tabular format and consists of 3445 visits with 159 diagnostic features collected, ranging from patient information to family history, to mobility scores and cerebral spinal fluid (CSF) markers...etc.

The heterogeneous tabular dataset is the most common type of dataset and is crucial for a myriad of critical applications, including medical data (27). *Tabular* data indicates the dataset is arranged in a table with data points as rows and features as columns. *Heterogeneous* indicates the dataset contains a variety of attribute types, including quantitative (numerical) and qualitative (categorical) variables. This is in contrast with homogeneous data which is characterized by a single data modality, such as image, text and audio data. In heterogeneous datasets, quantitative variables can be further divided into discrete or continuous while qualitative types can include ordinal and nominal types (28). For example, in PPMI clinical dataset, the discrete variables could include the number of *e4* alleles in the *APOE* gene, continuous variables could include the *height*, *weight* or the CSF (cerebral spinal fluid) *p-tau* levels. Nominal variables can include gender, ethnicity or the side of the body most affected by PD symptoms, Ordinal variables can include Epworth Sleepiness Scale Score, Tremor Score, and Geriatric Depression Scale Score (26).

Other than sharing the common challenges with analyzing homogeneous data, such as noise, impreciseness, missing data...etc., heterogeneous datasets present unique challenges. Specifically, the different attribute types lead to different ranges between variables and differences in sparsity between variables. For example, numerical features could be dense while categorical/nominal features could be very sparse after preprocessing by one-hot encoding. Further, the correlation amongst features of vary-

ing attribute types may be weaker than the spatial or semantic relationships as seen in image or text data. This is due to the arbitrariness of rows and columns, where switching their orders has no effect, whereas switching the orders of tokens in text data can dramatically change its meaning. Therefore, there is a need to uncover/exploit feature interactions without relying on positional/spatial information (29).

For the purpose of this paper, our goals are three-fold. We seek to analyze the PPMI clinical dataset to #1. validate that it contains sufficiently discriminant information for computer-based diagnostics, #2. to uncover important features for diagnostics based on algorithms used in the prior section, and #3. use the identified important features as candidates for our causal inference framework to examine potential causal relationships for molecular markers to PD symptomology and etiology.

3.2 Data preprocessing

The clinical dataset specifies 4 types of participants, including 2223 visits from PD participants, 1038 from healthy controls, and 184 from SWEDD participants, however, 0 recorded visits for the prodromal group. Data acquisition protocols and the PPMI database are available at `www.ppmi-info.org`. Before analyses, the dataset is arranged in a data matrix, where the rows represent one kind of object (i.e., patient state), and the columns represent another kind of object (i.e., clinical features), then 4 steps are taken to preprocess the clinical data.

Data exclusion The SWEDD cases are excluded due to their under-representation in data. Additionally, for the purpose of patient state classification, we have created 5 criteria for excluding features from analysis. *Firstly*, certain irrelevant features that do not pertain to diagnosis or patient state are excluded. This includes variables such as visit site, date or patient number. *Secondly*, features where the value only exists for PD or healthy patients are also excluded. This includes features like date of diagnosis and levodopa dose. *Thirdly*, features where the values indicate a clinical diagnosis are

discarded. This includes the most likely primary diagnosis ("primdiag") and comment for the most likely primary diagnosis ("othneuro"). *Fourthly*, features with mostly missing values and a few unique values are also discarded. This includes cases where molecular testing is below the limit of detection (e.g., "abeta_txt"). *Finally*, features that are redundant or older versions of new features are discarded (e.g., "fampd_new" is kept, over "fampd_old"). For a comprehensive list of the features excluded, refer to Appendix A.1.

Categorical variables preprocessing For categorical features, the ordinal features should be encoded by an ordinal encoder, while the nominal features are encoded with one-hot encoding. For this dataset, the nominal features include genotypes such as "APOE", "SNCA_rs356181", "SNCA_rs3910105" and "MAPT".

Numerical variables preprocessing For numerical features, they are all normalized to $0 \sim 1$ using min-max scaling. Some examples of numerical features in this dataset include: cerebral spinal fluid (CSF) *tau* levels, CSF hemoglobin values, CSF Aβ levels, and serum uric acid levels.

Data imputation Relative simple imputation methods are done for both categorical features and numerical features. For categorical features, all missing values are treated as one special category and replaced by a special value. For numerical features, all missing values of a variable are replaced by the mean of the existing values of that variable.

3.3 Predictive modelling: tree ensembles

3.3.1 Decision trees

The decision tree algorithm works by constructing a hierarchical structure that models decisions and their potential consequences through a tree-like structure where

each node contains conditional control statements (30).

For a classification problem with C classes and the original dataset X, during tree construction, at every iteration, the model recursively searches for the most optimal nodes for splits by employing an *exact greedy algorithm* whose goal is to minimize information entropy (H) or maximize the purity of the resulting subset (X) after the split (30). The entropy calculation is given by:

$$H(X) = \sum_{c \in C} -p(c) \log_2 p(c) \tag{3.1}$$

where C represents the set of classes in classes in the current dataset X, thus $p(c)$ represents the proportion of the number of elements of class c to the total number of elements in the subset X.

The algorithm enumerates over all unused features (F) and calculates the entropy of hypothetical subsets of data if the split were to happen with the said feature. Then, it calculates the difference in entropy between the subset before and the subset after the split on said feature. This difference is called "information gain" (IG), which represents the amount of uncertainty reduced in the subset by splitting on (30). The feature with the greatest information gain is picked as the node, and the iteration continues to find the next split until some pre-defined stopping criteria are met (30). IG is given by:

$$\begin{aligned} IG(X, F) &= H(X) - \sum_{s \in S} p(s) H(s) \\ &= H(X) - H(X|F) \end{aligned} \tag{3.2}$$

where $H(X)$ is the entropy (3.1) of subset X, S represents the resultant subset by splitting on feature F, thus $p(s)$ represents the proportion of the number of elements of s to the total number of elements in X.

After building the tree on the training set, at test time, the decision tree classifies new examples by traversing the entirety of the tree to arrive as a leaf node

– which will be the output (30).

Advantages of the decision tree algorithm include its interpretability, ability to visualize the decision-making process, being able to handle both numerical and categorical data, requiring little data preprocessing and having low cost of prediction in that it is logarithmic in the number of training data points. The disadvantages of the algorithm include being prone to overfitting, being unstable thus small variations in the dataset can completely change the tree structure, and certain concepts like XOR being hard to express (31). The first two problems can be overcome by tree ensembling.

3.3.2 Gradient boosting

Assuming a gradient boosting algorithm with M iterations, weak learners h_m are trained sequentially and then combined into a strong learner F_{m+1} to improve the performance of the ensemble model. Such as:

$$F_{m+1} = F_m + h_m(x_i) \tag{3.3}$$

where F_m represents the strong model at iteration m and h_m represent the weak learner at iteration m.

With each iteration, the error is calculated as some function of the prediction of the current strong model $F_m(x_i)$ and the ground-truth label y_i, then a new weak learner fits the error, also known as residual errors or negative gradients of the current strong model. At iteration m, the new weak learner fits the residual:

$$h_m(x_i) = y_i - F_m(x_i) \tag{3.4}$$

The general principle is F_{m+1} aims to correct the error of its predecessor F_m by adding a weaker learner fitted on the residual error to the ensemble (32).

In test time, the predictions of individual weak learners are combined to form a single output. In the case of classification problems, the class with the majority of the votes is used as the output (32). Assuming a dataset $\mathcal{D}$ with n examples and m features, $\mathcal{D} = \{x_i, y_i\}^n$ where $x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$ and assume M additive weak learners, the prediction for the output is:

$$\hat{y}_i = \phi(x_i) = \sum_{m=1}^{M} f_m(x_i), f_m \in \mathcal{F} \tag{3.5}$$

The computationally costly part of tree ensembling algorithms is finding the best split points when constructing new decision trees. There are two broad classes of algorithms for doing so. The first class includes the *pre-sorted* and *exact-greedy algorithm*, they operate by enumerating through all possible splits for feature values. This is inefficient and slow (33). The second class is the *histogram-based algorithm*, it works by separating feature values into bins and then construct feature histograms based on these bins. Comparatively, it is more efficient in training speed and memory consumption (34). This method is more efficient in both memory usage and training speed. Gradient-boosted tree ensemble variants tend to use different strategies to optimize finding the split-points.

3.3.3 XGBoost

XGBoost or eXtreme Gradient Boosting is a type of gradient-boosted tree ensemble algorithm. It has achieved state-of-the-art results on numerous machine learning competitions as well as a diverse range of practical tasks, including in the fields of energy consumption forecasting, failure detection for production lines, PCOS (polycystic ovary syndrome) diagnosis and seismic stability analysis...etc (35; 36; 37; 38).

Compared to traditional gradient boosting techniques, XGBoost or eXtreme Gradient Boosting made several innovations in algorithm and system design.

Algorithmically, XGBoost makes improvements on a regularized learning objective for gradient tree boosting algorithms, it also introduces an approximate algorithm as a more efficient alternative to the greedy algorithm for split finding along with modifications that allow for handling sparse input data (39). Systematically, XGBoost utilizes out-of-core computation and *cache-aware access* to enable parallel and distributed computation and thus faster model exploration (39). All XGBoost experiments conducted for this study uses the Python API of the official XGBoost implementation from: `https://github.com/dmlc/xgboost`.

The criteria for evaluating potential splits in gradient-boosted tree construction are as follows:

$$\mathcal{L}_{\text{split}} = \frac{1}{2}[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L\ h_i+\lambda}} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}] - \gamma \tag{3.6}$$

where I represents the instance set of a leaf, I_L, I_R represents the instance sets of left and right nodes after a split, g and h represent first and second-order gradient statistics respectively.

3.4 Causal inference

After the identification of features most important for classification according to XGBoost, we divide the features according to their varying levels of biology (i.e., molecular level, cellular level, tissue level, organ and symptom level), and attempt to establish causal relationships from lower levels of pathology (e.g., molecular) to higher levels of pathology (e.g., symptoms). To achieve this, we employ Sharma and Kiciman's 4-step process for causal analyses which are further described below (40).

Model the causal question This step involves the creation of a causal graph to represent prior knowledge and to explicitly structure one's causal assumptions regarding a subset of variables (41). For the problem of identifying disease-causing

mechanisms, we make 2 broad assumptions. Firstly, biology is hierarchical, changes on the molecular level affect changes on the cellular level, then the tissue level... and eventually are exhibited as symptoms (42). The levels of organization in biology are illustrated at 3.1 and the causal graph for disease pathogenesis is illustrated as 4.3. Secondly, feature importance calculation coupled with ablation studies with high-performing classifiers is a principled way of identifying and isolating variables of interest.

Notably, the rest of the variables not represented in the causal graph are assumed as potential confounders (40). Confounders are variables that influence both the treatment and outcome variable, but it does not affect the outcome through its impact on treatment (43). Confounding variables are known to cause spurious association and are threats to internal validity, thus their effects must be suppressed in order to establish causality (44). If X represent the treatment and Y represent the outcome variable, X and Y are not confounded if and only if:

$$P(y|\text{do}(x) = p(y|x) \tag{3.7}$$

where $P(y|\text{do}(x)$ denotes the probability of outcome y under the hypothetical treatment of x.

Alternatively, if there is a variable confounding the relationship between X and Y, we have:

$$P(y|\text{do}(x) \neq p(y|x) \tag{3.8}$$

Identify the causal estimand depending on the causal graph, various graph-based methods can be used to identify the causal effect. Sharma & Kiciman utilize graph-based criteria and do-calculus to find the appropriate identification criteria, including the back-door criterion, the front-door criterion, and mediation...etc (45; 40).

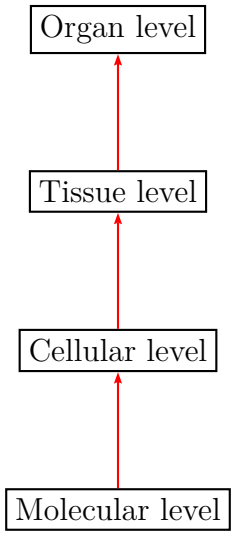


Figure 3.1: Levels of organization in biology, with the highest level at the top and the lowest level at the bottom.

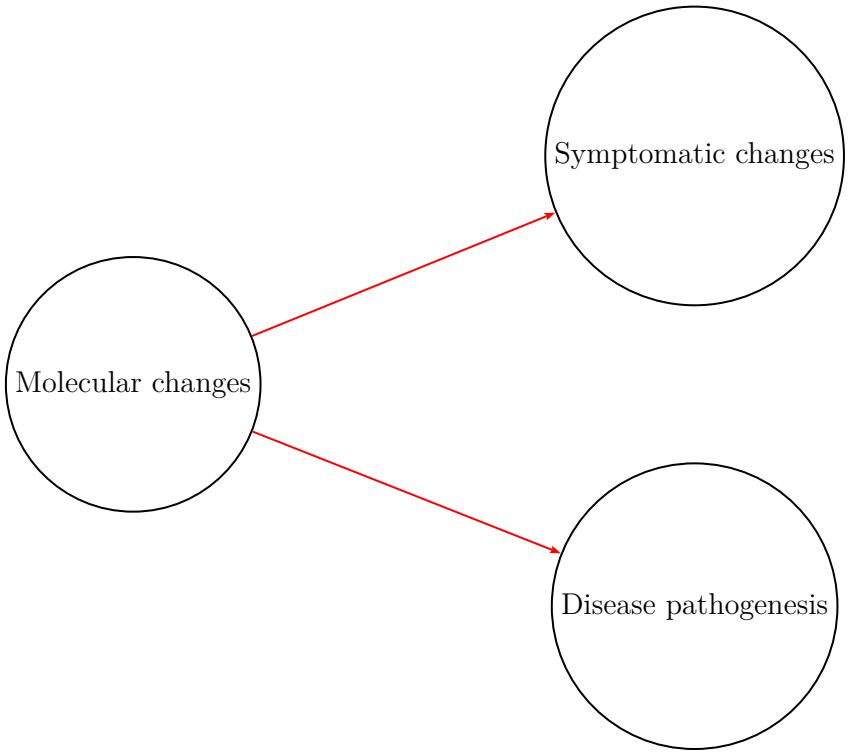


Figure 3.2: Causal graph for disease pathogenesis.

The rationale is, in the case of the existence of confounders, an adjustment formula will need to be utilized to adjust for all confounding factors C in order to attain unbiased estimates of the causal effect or $P(y|\text{do}(x)$ (43). This is expressed as:

$$P(y|\text{do}(x) = \sum_{c} p(y|x, c)p(c) \tag{3.9}$$

Similarly, the backdoor criterion involves conditioning on confounders to intercept all paths between X and Y with an arrow towards X. The intuition is, if these paths are blocked, then observed causal effects reflect $P(y|\text{do}(x)$ (43). Beyond adjustment and backdoor criterion, Pearl's do-calculus provides other methods for an unbiased estimation of $P(y|\text{do}(x)$ (45).

Estimate the causal effect employ statistical estimators to quantify the identified estimand. Examples of estimators for the outcome model or ATE include linear regression and generalized linear models (43). We will use the average treatment effect (ATE) estimand, which is given by:

$$\text{ATE} = \mathbb{E}[Y(1) = Y(0)] \tag{3.10}$$

where $\mathbb{E}$ represents the expected value, $Y(1)$ represents the outcome variable for the treatment group and $Y(0)$ represents the outcome variable for the control group.

For this study, we employ the backdoor criterion with linear regression for ATE estimation.

Refute the obtained estimate different from supervised learning methods such as decision trees or methods of gradient-boosted tree ensembles where model validation can be done with a held-out set or by k-fold cross-validation, causal tasks often do not have access to ground-truth labels. Validation of causal assumptions is usually

done with a series of robustness and sensitivity checks. A general principle for these tasks including 3 tasks

- **Randomly permute the treatments**, therefore destroying the existing causal effect and measure the probability of seeing an effect as strong as the previous causal effect. To pass this test, the probability should be < 0.05.
- **Randomly permute the outcome**, therefore destroying the existing causal effect and measure the probability of seeing no causal effect. To pass this test, the probability should be 0 or close to 0.
- **Randomly generate confounders for both the treatment and the outcome**, the new confounder variables are sampled from a normal distribution, then we measure the probability of the original estimate being within the distribution of the now permuted estimates. To pass this test, there should be 0 difference between the original and permuted effect.

All validation tests are performed using the DoWhy library via the `causalwizard.app` (40).

Chapter 4: Results

4.1 XGBoost classification results

Experimental classification results from 5-fold cross-validation of the XGBoost with default hyperparameters indicate a near-perfect performance of a mean accuracy of 97.80% and a standard deviation of 0.0094 as seen in 4.1 (39). Meaning, given the 979 samples, XGBoost incorrectly classifies 21 samples.

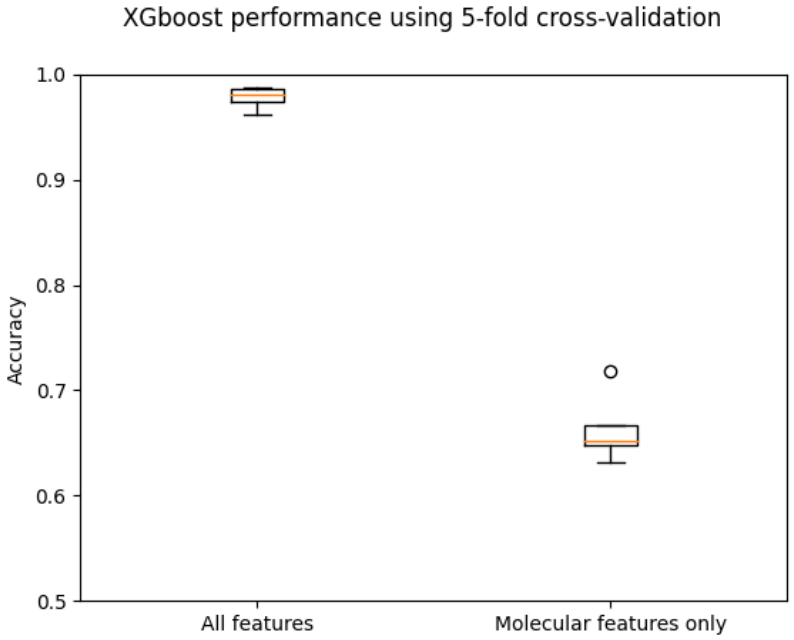
4.2 Ablation studies

When we isolate only the molecular features for classification using XGBoost, its performance as validated by 5-fold cross-validation has a significant drop with a mean accuracy of 66.30% and a standard deviation of 0.029 (4.1).

4.3 Feature importance

During the construction of decision trees by XGBoost, the motor-symptom feature "rigidity_0" has the highest feature importance at 29.32%, while a non-motor symptom "upsit_cat" (University of Pennsylvania Smell Identification Test) has the 2nd higher feature importance with a score of 21.91%. This indicates their high discriminative power for classifying PD and healthy subjects. No other feature has a feature importance greater than 10% (4.2). Alternatively, the "age" feature has 1 of the lowest feature importance, this makes sense as the PPMI clinical dataset controls for age of its subjects, and there is little difference in the average age between healthy and PD subjects.

When all the other features are removed in favor of molecular features, SNCA (α-synuclein) genotype-related markers have the highest feature importance, including "SNCA_rs3910105_0" with 10.96% and "SNCA_rs356181_0" with 10.87% fea-



(a) A box and whisker plot is a comparison of the performance of XGBoost on all the features vs. using only the molecular features from the PPMI clinical dataset.

	Mean accuracy, STD
XGBoost (**All features**)	0.9779, 0.0094
XGBoost (**Only molecular features**)	0.6629, 0.0295

(b) Mean accuracy and standard deviation of XGBoost classification results on all features vs. only molecular features

Figure 4.1: Ablation study results

ture importance respectively. No other molecular feature has a feature importance greater than 10% in the ablation case (4.2). Among the molecular features with the lowest importance are "abeta" (Aβ) and "ab_asyn" (ratio of Aβ to α-synuclein) suggesting has low discriminative power for Aβ-related features for classifying healthy and PD subjects in this case.

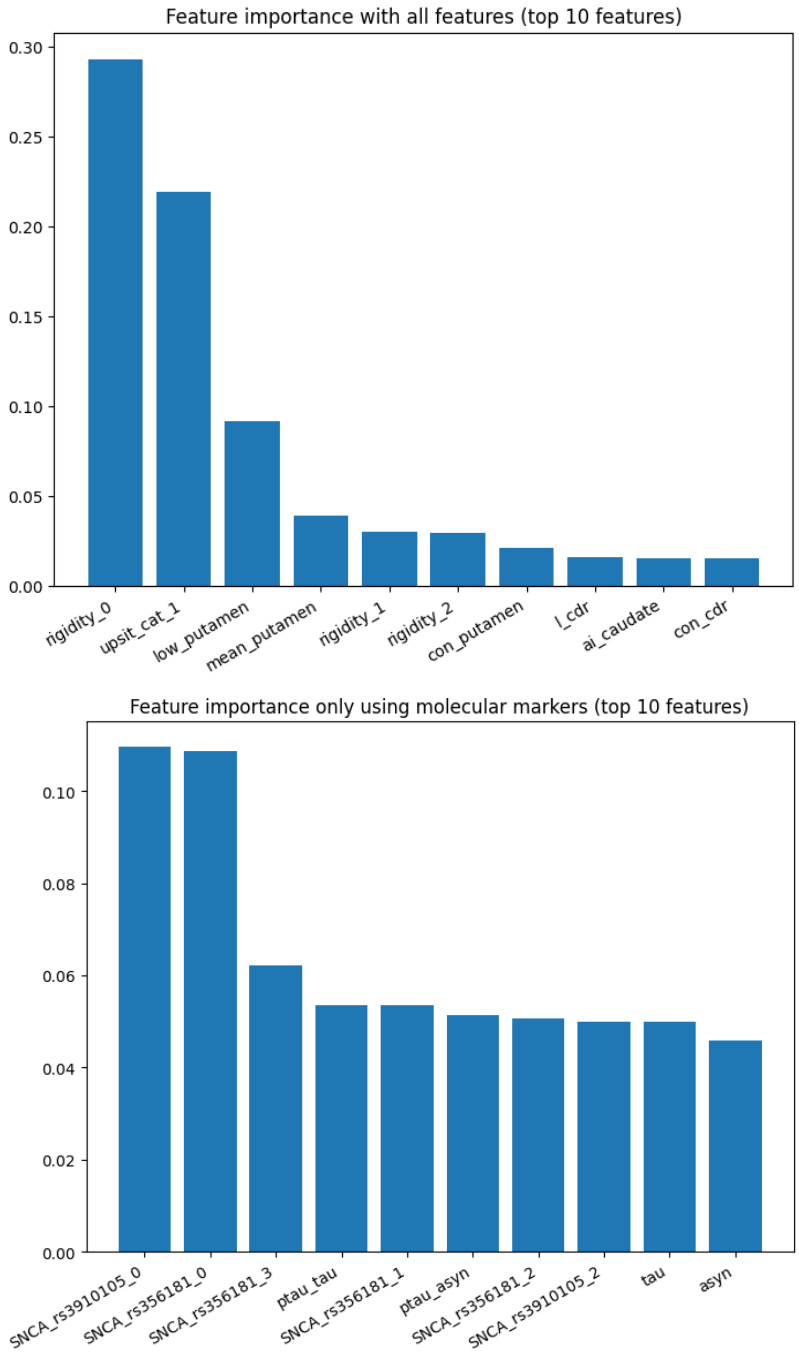


Figure 4.2: Top 10 most important features for XGBoost

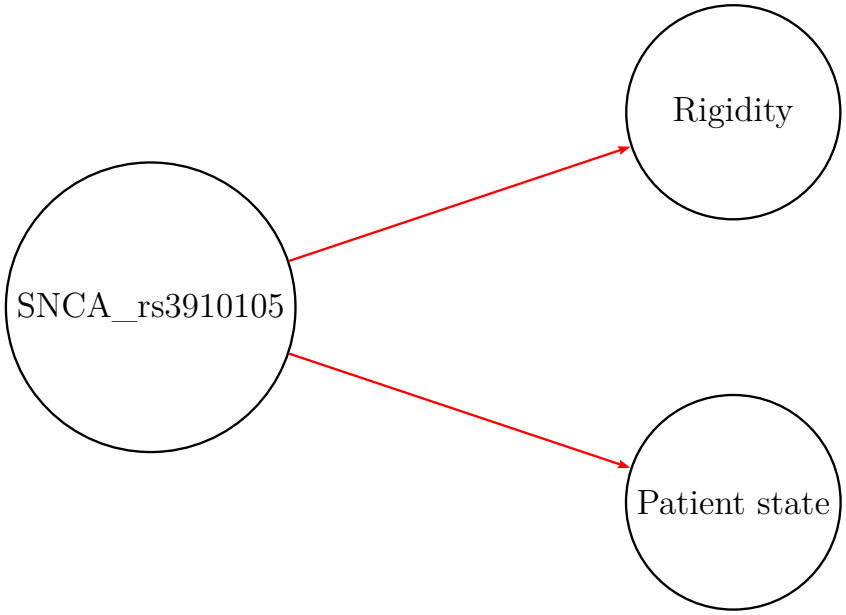


Figure 4.3: Causal graph for disease pathogenesis as informed by features with the highest feature importance in XGBoost classification.

4.4 Establishing causation

With the isolation of the most important molecular and clinical or symptom marker and coupled with the prior knowledge of hierarchies of biology, we attempt to examine whether there are causal relationships between "SNCA_rs3910105_0" and "rigidity_0", and "SNCA_rs3910105_0" and the patient diagnosis (3.1, 4.3).

First, we create the causal graph (4.3).

Molecular → Symptom : with the molecular variable "SNCA_rs3910105" as the treatment, and the clinical/symptom variable "rigidity" as outcome, we perform linear regression with back-door criterion to estimate the ATE (average treatment effect) the treatment on outcome. We find the ATE = 0.52, meaning when treatment SNCA_rs3910105 > 0, the value of outcome rigidity is on average increased by 0.52 units compared to when SNCA_rs3910105 ≤ 0 (4.1).

This result passes all the validation & robustness tests mentioned in 3.4, as

Causal relationship	Average treatment effect (ATE)	Validation tests
Molecular $\rightarrow$ Symptom	0.52 (95% Confidence Interval $[0.27, 0.77]$)	Passed (B.1)
Molecular $\rightarrow$ Patient state	0.14 (95% Confidence Interval $[0.1, 0.17]$)	Passed (B.2)
Symptom $\rightarrow$ Patient state	0.57 (95% Confidence Interval $[0.54, 0.59]$)	Passed (B.3)

Table 4.1: Causal inference results

shown in appendix B.1.

Molecular $\rightarrow$ Patient state with the molecular variable "SNCA_rs3910105" as the treatment, and the patient state (healthy or PD) as outcome, we perform linear regression with back-door criterion to estimate the ATE the treatment on outcome. We find the ATE = 0.14, meaning when treatment SNCA_rs3910105 > 0, the probability of patient state being PD is on average increased by 0.14 compared to when SNCA_rs3910105 ≤ 0 (4.1).

This result passes all the validation & robustness tests mentioned in 3.4, as shown in appendix B.2.

Symptom $\rightarrow$ Patient state with the clinical/symptom variable "rigidity" as the treatment, and the patient state (healthy or PD) as outcome, we perform linear regression with back-door criterion to estimate the ATE the treatment on outcome. We find the ATE = 0.57, meaning when treatment rigidity > 0, the probability of the patient state being PD is on average increased by 0.57 compared to when rigidity ≤ 0 (4.1).

This result passes all the validation & robustness tests mentioned in 3.4, as shown in appendix B.3.

Chapter 5: Conclusion & discussions

In this study, we sought to explore the pathogenesis of PD – an extremely heterogeneous disease, both in its etiology and presentation, of which there are no effective treatments for today.

We investigated the PPMI clinical dataset which contains a variety of types of features including demographic, genetic & molecular, imaging, symptomatic and the corresponding patient state.

We have shown XGBoost with default hyperparameters can achieve high predictive accuracy for the classification of healthy vs. PD on the PPMI clinical dataset. This indicates the features collected carry sufficient discriminative power to delineate patient states.

Further, by using the feature importance score in XGBoost and ablation methods, we can identify and isolate a few features of high importance that encompass varying levels of biology, and use them to inform the structure of causal graphs in combination with some prior knowledge. We found the genotype feature SNCA_rs3910105 and the clinical/symptom feature rigidity to have higher feature importance and hypothesized the former feature having a causal relationship with the latter feature.

In the end, we were able to establish the causal relationship and quantify the ATE of a series of treatment variables on outcome variables, including SNCA_rs3910105 $\rightarrow$ rigidity, SNCA_rs3910105 $\rightarrow$ patient state, and rigidity $\rightarrow$ patient state. Our findings indicate that SNCA_rs3910105 causes the symptom of rigidity – which is a hallmark of PD, we also further validated the viability of using rigidity as a diagnostic marker for PD, that the presence of rigidity beyond a certain scale has a causal effect on the diagnosis of PD. Our results also reaffirm the usefulness of rigidity as a criterion in clinical practices (24).

The main contribution of this study is two folds. Firstly, we initially use a correlative method in gradient-boosted decision trees to verify there exists sufficient discriminative information in the dataset to accurately predict healthy vs. PD state, then quantifying & isolating the most important features for further investigation in a causal inference framework. One can view the objective function of minimizing information entropy is used as a proxy to identify potential causal variables, which allows for the automatic translation of information in the dataset to causal graphs. Secondly, the findings indicate a single genotype marker SNCA_rs391010 as a causal mechanism for rigidity and PD diagnosis. This supports the current hypothesis of the role of aggregation of misfolded $\alpha-$synuclein proteins in PD pathogenesis (15).

Ultimately, the establishment of causal relationships poses many advantages over correlative relationships. Especially in high-stake sectors such as healthcare, scientists & clinicians can devote more resources addressing the causal variable knowing the observed effects are likely not due to confounders or spurious correlation – which if research & development efforts were based on, may not lead to medical interventions that confer benefits to patients.

Appendix A: PPMI data preprocessing

	Features that fit the criteria
Criteria 1	SITE, PATNO, EVENT_ID, YEAR, visit_date
Criteria 2	agediag, duration, ageonset DOMSIDE, LEDD, MSEADLG, PD_MED_USE symptom1, symptom2, symptom3, symptom4 symptom5, symptom5_comment, symptom6 hy, hy_on, nhy, nhy_on updrs1_score, updrs2_score, updrs3_score, updrs3_score_on updrs4_score, updrs_totscore, updrs_totscore_on ST_startdate, ST_year1, ST_year2, ST_year3, ST_year4, ST_year5
Criteria 3	othneuro, primdiag
Criteria 4	abeta_txt, tau_txt, ptau_txt
Criteria 5	fampd_old, td_pigd_old, td_pigd_old_on

Table A.1: 5 criteria for feature exclusion for the PPMI clinical dataset to enable for unbiased & efficient analyses. Rationale for data exclusion found at 3.2. PPMI dataset dictionary, along with variable and schema definitions can be found at `https://www.ppmi-info.org/access-data-specimens/data-dictionary`

Appendix B: Causal inference validation tests

Validation

Several tests were applied to assess the significance and robustness of the results.

Method	Result	Desired result	Test
Bootstrap outcome permutation (statistical significance test) n=100 Generates n datasets with randomly permuted **outcomes**, destroying causal effect. Measures probability of seeing a result as strong as the original causal effect, when it doesn't exist. Read more	p = [0, 0.01]	p <= 0.05	Pass
Replace treatment with placebo n=100 Generates n datasets with randomly permuted **treatments**, destroying causal effect. Evaluates probability than a causal effect of zero is in the distribution of n effect estimates. Read more	Original effect: 0.52 Placebo effect: -0.01 p=0.88	Effect **becomes zero**	Pass
Add Random common cause (confounder) of both treatment and outcome n=100 Generates n datasets with an additional normally-distributed random variable, representing an observed (or unobserved) confounder. Significance test determines probability that original estimate lies within the distribution of n modified estimates. Read more	Original effect: 0.52 Modified effect: 0.52 p=0.98	Effect **unchanged**	Pass

Figure B.1: Validation tests passed for establishing causal relationship between SNCA_rs3910105 and rigidity.

Validation

Several tests were applied to assess the significance and robustness of the results.

Method	Result	Desired result	Test
Bootstrap outcome permutation (statistical significance test) n=100 Generates n datasets with randomly permuted **outcomes**, destroying causal effect. Measures probability of seeing a result as strong as the original causal effect, when it doesn't exist. Read more	p = [0, 0.01]	p <= 0.05	*Pass*
Replace treatment with placebo n=100 Generates n datasets with randomly permuted **treatments**, destroying causal effect. Evaluates probability than a causal effect of zero is in the distribution of n effect estimates. Read more	Original effect: 0.14 Placebo effect: 0 p=0.98	Effect **becomes zero**	*Pass*
Add Random common cause (confounder) of both treatment and outcome n=100 Generates n datasets with an additional normally-distributed random variable, representing an observed (or unobserved) confounder. Significance test determines probability that original estimate lies within the distribution of n modified estimates. Read more	Original effect: 0.14 Modified effect: 0.14 p=0.96	Effect **unchanged**	*Pass*

Figure B.2: Validation tests passed for establishing causal relationship between SNCA_rs3910105 and patient state (healthy or PD).

Validation

Several tests were applied to assess the significance and robustness of the results.

Method	Result	Desired result	Test
Bootstrap outcome permutation (statistical significance test) n=100 Generates n datasets with randomly permuted **outcomes**, destroying causal effect. Measures probability of seeing a result as strong as the original causal effect, when it doesn't exist. Read more	p = [0, 0.01]	p <= 0.05	*Pass*
Replace treatment with placebo n=100 Generates n datasets with randomly permuted **treatments**, destroying causal effect. Evaluates probability than a causal effect of zero is in the distribution of n effect estimates. Read more	Original effect: 0.57 Placebo effect: 0 p=0.92	Effect **becomes zero**	*Pass*
Add Random common cause (confounder) of both treatment and outcome n=100 Generates n datasets with an additional normally-distributed random variable, representing an observed (or unobserved) confounder. Significance test determines probability that original estimate lies within the distribution of n modified estimates. Read more	Original effect: 0.57 Modified effect: 0.57 p=0.98	Effect **unchanged**	*Pass*

Figure B.3: Validation tests passed for establishing causal relationship between rigidity and patient state (healthy or PD).

Works Cited

[1] Zoltan Hermann, Hedvig Horváth, and Attila Lindner. Answering causal questions using observational data-achievements of the 2021 nobel laureates in economics. *Financial and Economic Review*, 21(1):141–163, 2022.

[2] Kenneth Stanley. Design of randomized controlled trials. *Circulation*, 115(9):1164–1169, 2007.

[3] Petra E Todd and Kenneth I Wolpin. Ex ante evaluation of social programs. *Annales d'Economie et de Statistique*, pages 263–291, 2008.

[4] Petra E Todd and Kenneth I Wolpin. Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American economic review*, 96(5):1384–1417, 2006.

[5] Hakan Seckinelgin. *Politics of Global Aids*. Springer, 2017.

[6] James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007.

[7] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

[8] Amar Bhide, Prakesh S Shah, and Ganesh Acharya. A simplified guide to randomized controlled trials. *Acta obstetricia et gynecologica Scandinavica*, 97(4):380–387, 2018.

[9] Fernando P Polack, Stephen J Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L Perez, Gonzalo Pérez Marc, Edson D Moreira, Cristiano Zerbini, et al. Safety and efficacy of the bnt162b2

mrna covid-19 vaccine. *New England journal of medicine*, 383(27):2603–2615, 2020.

[10] Oliver J Watson, Gregory Barnsley, Jaspreet Toor, Alexandra B Hogan, Peter Winskill, and Azra C Ghani. Global impact of the first year of covid-19 vaccination: a mathematical modelling study. *The Lancet Infectious Diseases*, 22(9):1293–1302, 2022.

[11] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W Green, and Cate D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2):155–161, 2007.

[12] Stephen H Bell and Laura R Peck. Obstacles to and limitations of social experiments: 15 false alarms. *Abt thought leadership paper, Abt Associates*, 2012.

[13] David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

[14] Bastiaan R Bloem, Michael S Okun, and Christine Klein. Parkinson's disease. *The Lancet*, 397(10291):2284–2303, 2021.

[15] Panchanan Maiti, Jayeeta Manna, and Gary L Dunbar. Current understanding of the molecular mechanisms in parkinson's disease: Targets for potential treatments. *Translational neurodegeneration*, 6:1–35, 2017.

[16] Lorraine V Kalia and Anthony E Lang. Parkinson's disease. *The Lancet*, 386(9996):896–912, 2015.

[17] Eduardo Tolosa, Alicia Garrido, Sonja W Scholz, and Werner Poewe. Challenges in the diagnosis of parkinson's disease. *The Lancet Neurology*, 20(5):385–397, 2021.

[18] E Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. The emerging evidence of the parkinson pandemic. *Journal of Parkinson's disease*, 8(s1):S3–S8, 2018.

[19] Wenya Yang, Jamie L Hamilton, Catherine Kopil, James C Beck, Caroline M Tanner, Roger L Albin, E Ray Dorsey, Nabila Dahodwala, Inna Cintina, Paul Hogan, et al. Current and projected future economic burden of parkinson's disease in the us. *npj Parkinson's Disease*, 6(1):15, 2020.

[20] Garrett E Alexander. Biology of parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. *Dialogues in clinical neuroscience*, 6(3):259–280, 2004.

[21] C Colosimo, AJ Hughes, L Kilford, and AJ Lees. Lewy body cortical involvement may not always predict dementia in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(7):852–856, 2003.

[22] WR Gibb, CQ Mountjoy, DM Mann, and AJ Lees. A pathological study of the association between lewy body disease and alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 52(6):701–708, 1989.

[23] CE Clarke, S Patel, N Ives, CE Rick, R Woolley, K Wheatley, MF Walker, S Zhu, R Kandiyali, G Yao, et al. Uk parkinson's disease society brain bank diagnostic criteria. *NIHR Journals Library*, 2016.

[24] Ronald B Postuma, Daniela Berg, Matthew Stern, Werner Poewe, C Warren Olanow, Wolfgang Oertel, José Obeso, Kenneth Marek, Irene Litvan, Anthony E Lang, et al. Mds clinical diagnostic criteria for parkinson's disease. *Movement disorders*, 30(12):1591–1601, 2015.

[25] Ronald B Postuma, Werner Poewe, Irene Litvan, Simon Lewis, Anthony E Lang, Glenda Halliday, Christopher G Goetz, Piu Chan, Elizabeth Slow, Klaus Seppi,

et al. Validation of the mds clinical diagnostic criteria for parkinson's disease. *Movement Disorders*, 33(10):1601–1608, 2018.

[26] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M Tanner, John Q Trojanowski, et al. The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477, 2018.

[27] Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pages 341–354. PMLR, 2020.

[28] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[29] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[30] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[31] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[32] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[33] Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. Sliq: A fast scalable classifier for data mining. In *Advances in Database Technology—EDBT'96: 5th International Conference on Extending Database Technology Avignon, France, March 25–29, 1996 Proceedings 5*, pages 18–32. Springer, 1996.

[34] Ping Li, Qiang Wu, and Christopher Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in neural information processing systems*, 20, 2007.

[35] Abnash Bassi, Anika Shenoy, Arjun Sharma, Hanna Sigurdson, Connor Glossop, and Jonathan H Chan. Building energy consumption forecasting: A comparison of gradient boosting models. In *The 12th International Conference on Advances in Information Technology*, pages 1–9, 2021.

[36] Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, and Rizwan Hasan. Improved sampling and feature selection to support extreme gradient boosting for pcos diagnosis. In *2021 IEEE 11th annual computing and communication workshop and conference (CCWC)*, pages 1046–1050. IEEE, 2021.

[37] Ankita Mangal and Nishant Kumar. Using big data to enhance the bosch production line performance: A kaggle challenge. In *2016 IEEE international conference on big data (big data)*, pages 2029–2035. IEEE, 2016.

[38] Luqi Wang, Jiahao Wu, Wengang Zhang, Lin Wang, and Wei Cui. Efficient seismic stability analysis of embankment slopes subjected to water level changes using gradient boosting algorithms. *Frontiers in Earth Science*, 9:807317, 2021.

[39] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[40] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[41] Judea Pearl. *Causality*. Cambridge university press, 2009.

[42] Markus I. Eronen and Daniel Stephen Brooks. Levels of Organization in Biology. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[43] Judea Pearl et al. Models, reasoning and inference, second edition. *Cambridge, UK: CambridgeUniversityPress*, 2009.

[44] Christine M Anderson-Cook. Experimental and quasi-experimental designs for generalized causal inference, 2005.

[45] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.