A Causal Study of Physical Health's Role in Heart Diseases

Asritha Pidikiti Sanjna Chippalaturthi Mounika Boggada

1 Abstract

Heart diseases is the leading cause of death around the globe, and the importance of learning how physical activity leads to decreased risk in such cases is necessary. The project, in detail, underlines the directed acyclic graph (DAG), constructed from the entire dataset, that explores issues regarding the association between physical activity and heart disease causation. The DAG focused on identifying core variables under the guidance of established directions of causation, putting in evidence some confounders like *BMI*, *Smoking*, *PhysicalHealth*, etc. Using the S-learner framework with Random Forests, we estimated both the Average Treatment Effect (ATE) and Heterogeneous Treatment Effects (HTE).

Under the ATE, it was found that the probabilities of occurrence of heart disease among physically active individuals were lower than those with inactivity by about 3.5%. Under HTE, the analysis revealed varied treatment effects across individual levels. Validation techniques, including Estimated Root Mean Squared Error (ERMSE), confirmed the robustness of the findings, while SHAP analysis showed that predictors such as *PhysicalHealth*, *BMI*, and *Smoking* were identified as key features. Hence, these results shed light on causal pathways and provide solid evidence in favor of public health policies for physical activity as a way of combating heart disease.

2 Introduction

Motivation

Cardiovascular conditions, such as heart disease, are the main causes of deaths globally, causing millions of people to die each year. Apart from the severe health effects, the economic losses from cardiovascular diseases are costly, making it a worldwide issue. Regular physical activity has been widely-approved as one of the means of preventing heart disease, yet the exact cause is not sufficiently researched.

Research Objective

This project is an innovative effort to detect and deal with the overload in heart disease by examining the accepted effect of physical activity using modern approaches. While the therapeutic potential of physical activity is universally recognized, a definite model of the cause-and-effect relationship between it and heart disease is scarce due to the complexities.

Among the methods that were used for the research of the causal relationship between physical activity and heart disease, we started with causal graph theory and heterogeneous treatment effect estimation. We have been able to identify statistically significant relationships in physical activity with health by this approach that also helped the control confounding variables and analyses of the individual effects of physical activity with vivid results. Therefore, we got the necessary data for the nuanced insights and evidence-based information that helped in the designing of the public health policies that are meant to minimize the risks of heart disease.

Approach and Contributions

We utilized a method to assess two Average Treatment Effects (ATE) and Heterogeneous Treatment Effects (HTE) of physical activity in heart disease. The crucial notions are as follows:

- Deciding on the Directed Acyclic Graphs (DAGs) strategy to discover and handle confounders, therefore, sustainable and reliable causal estimates.
- Adopting specific approaches-based on backdoor adjustment, propensity score matching (PSM), and the inverse probability of treatment weighting (IPTW) to quantile causal effects.
- Analyzing HTE to estimate the effects of individual treatment and identify variations in the effect of physical activity on heart disease risk based on demographic and health-related predictors.

Doing research on the interaction of physical activities with misleading data, this venture gives to policymakers valuable discoveries. Specific data suggest the new "evidential" public health models, particularly the heart disease risk reduction, bring health advancements to all or society.

3 Related Work

DAGs are very popular for encoding causal structures and helping identify relationships from observational data. This technique has even been employed to study risk for outcomes such as heart disease in healthcare. Greenland et al. (2009) and Hernán Robins (2020) emphasize the necessity of confounder identification for an unconfounded estimate of causal effects. This study used the PC algorithm, a constraint-based method of DAG construction that relies on conditional independence tests. Further work by Colombo and Maathuis (2014) showed that this algorithm is robust in high-dimensional data. We used the PC algorithm in a way to find some of the important confounders, like Smoking, BMI, and PhysicalHealth. The confounders affect both treatment (physical activity) and the outcome (ground for heart disease) and allow valid causation estimation using the backdoor adjustment criterion. Moreover, domain knowledge was incorporated into the learned structure of the DAG to ensure the structure met the known causal relationship, which improves existing methods.

An average treatment effect (ATE) is an estimation that focuses on causal inference-it makes possible an unbiased estimation of the causal effect of treatment on the material. One of the simplest backdoor adjustments by Pearl is the identification of confounders that block one from backdoor paths from treatment XX to outcome YY. By conditioning on such confounders, one can bring about unbiased estimates of causal effects. Using propensity score methods but with clear emphasis on Inverse Probability of Treatment Weighting (IPTW), backdoor adjustments further reinterpret observations as weights which then transform the original study population into a pseudo-population where treatment assignment differs from confounders. Cobalt balance is crucial for IPTW-based ATE estimate according to Austin and Stuart, who would have it assessed by standard criteria, for example, the Standardized Mean Difference (SMD). This paper discusses IPTW in terms of estimating propensity scores, checking covariate balance, and measuring ATE uncertainty via bootstrap confidence intervals.

The estimation of heterogeneous treatment effects (HTE) is becoming increasingly popular for estimating treatment effects on different individuals or even subgroups, thus providing the necessary information for personal interventions. HTE would reach out to the specific population concerning its own effects, while on the wider scale ATE would refer to population-level averages. Meta-learners, as S-learners and T-learners, are very flexible frameworks for integration of different machine learning models into causal inference. According to Shalit et al. (2017), metrics like ERMSE determined differences in predicted treatment effects between treatment and control groups, thus facilitating the evaluation of model robustness. Ling et al. (2022) further showed that S-learners with Random forests are very effective in estimating reliable HTE, thereby balancing interpretability and accuracy. Study uses these methods, conducts a SHAP analysis to measure feature importance and gives actionable recommendations for personalized health interventions.

This study ties the learning of causal structures from DAGs with both the ATE and HTE estimation models, thus showing the power of integrated approaches in healthcare. Using a combination of the PC algorithm, IPTW, and even more advanced machine learning methodologies, the present study gives very robust and interpretable answers concerning the causal pathways between physical activity and heart disease for population-level as well as individualized interventions.

4 Formal Problem Description

Problem Statement

The main idea of this investigation is to come up with the causal effect of physical activity on heart disease using observational data that come from better health, National Health Information Database (NHID), and smoking. In a more detailed manner, we want to yield information on not only the Average Treatment Effect (ATE) but also Heterogeneous Treatment Effects (HTE) which will uncover both the general influence of physical activity on human health and its variation within distinct subgroups.

Notation and Definitions

- **Data:** Let the dataset consist of n individuals, where each individual is represented by:

  $$(X_i, A_i, Y_i), \quad i = 1, 2, \ldots, n$$

  where:

  X_i: Vector of covariates (e.g., age, BMI, smoking, physical health).

  A_i: Treatment variable ($A_i = 1$ if the individual is physically active, $A_i = 0$ otherwise).

  Y_i: Outcome variable ($Y_i = 1$ if the individual has heart disease, $Y_i = 0$ otherwise).
- **Objective:**
  - **Average Treatment Effect (ATE):**

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

    where $Y(1)$ and $Y(0)$ denote the potential outcomes if an individual is treated (physically active) or untreated (physically inactive), respectively.
  - **Heterogeneous Treatment Effects (HTE):**

    $$\text{HTE}(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$$

    which quantifies how the treatment effect varies across subgroups defined by covariates X.
- **Assumptions:**
  - **Positivity:** $0 < P(A = 1 \mid X) < 1$, ensuring that all individuals have a non-zero probability of being treated or untreated.
  - **Conditional Ignorability:** $(Y(1), Y(0)) \perp\!\!\!\perp A \mid X$, meaning that treatment assignment is independent of the potential outcomes given the covariates.
  - **Consistency:** Observed outcomes correspond to the potential outcome under the assigned treatment.

Methodological Framework

- **Modeling the Causal Structure:**
  - A Directed Acyclic Graph (DAG) was established to illustrate the causal relationships among variables and locate confounders
  - ATE analysis included confounders to block backdoor paths and get accurate estimates.
- **Treatment Effect Estimation:**
  - **Back-door Adjustment:** Controlled for confounders by using the regression model of propensity scores.
  - **Propensity Score Methods:** The Propensity Score Matching (PSM) and the Inverse Probability of Treatment Weighting (IPTW) were employed to bring the covariates of the treatment groups in line with each other.
- **Heterogeneity Analysis:**
  - Subgroup analyses were conducted to evaluate how the effect of physical activity varied based on covariates like age, BMI, and smoking status.

Our Aim to Solve

- **Quantify the Effect:** Accurately measure the causal effect of physical activity on heart disease, controlling for confounders.
- **Policy Recommendations:** Develop evidence-based guidelines to promote physical activity as a targeted intervention for reducing heart disease risk.

5 Solutions

5.1 DAG Analysis

Data Preparation

The analysis began with all variables from the dataset as explored during EDA, ensuring that the DAG reflected the full scope of the dataset without excluding potentially influential factors. These variables included demographic attributes (e.g., *AgeCategory*, *Sex*, *Race*), health indicators (e.g., *BMI*, *PhysicalHealth*, *GeneralHealth*), and lifestyle habits (e.g., *Smoking*, *AlcoholDrinking*, *SleepTime*, *PhysicalActivity*). All variables were processed and incorporated into the DAG to capture their relationships (both direct and indirect) with the treatment (*PhysicalActivity*) and outcome (*HeartDisease*) variables. Though this added complexity, it ensured that no relevant relationships were overlooked during the initial DAG construction. Then, we refined the DAG to focus on simplifying the graph for clarity.

DAG Initialization and Refinement

The DAG was constructed as a directed graph using the *NetworkX* library, with nodes representing all dataset variables. Initial causal relationships were inferred using the PC Algorithm based on variable correlations. However, the algorithm considered relations between all variables, which resulted in an overly complicated DAG with dense interconnections.

To simplify this, the analysis focused specifically on relationships involving the treatment (*PhysicalActivity*) and outcome (*HeartDisease*) variables. This ensured that only relevant causal paths were considered.

Identifying Confounders

Confounders were defined as variables influencing both *PhysicalActivity* and *HeartDisease* based on the DAG structure. The identified confounders were:

- *BMI*
- *Smoking*
- *PhysicalHealth*

These variables were deemed critical for blocking back-door paths and ensuring unbiased causal effect estimation.

Challenges Addressed

- **Overly Complex DAG:** The PC Algorithm initially created a very complex DAG with excessive interconnections. Focusing on treatment-outcome relationships simplified the graph while retaining relevance.
- **Balancing Completeness and Simplicity:** Including all variables ensured no information was omitted, but careful refinement focused on simplifying the analysis to relationships relevant to the research question.

5.2 Back-Door Adjustment Using Causal Diagrams

Using the DoWhy, framework we conduct the back-door adjustment for estimating the causal effect of physical activity on heart disease. The treatment (AAA, where A=1A = 1A=1 for physically active and A=0A = 0A=0 for inactive), outcome (YYY, where Y=1Y = 1Y=1 for heart disease and Y=0Y = 0Y=0 otherwise), and confounders ((X, including age, BMI, smoking, and physical health) were introduced in the causal model. The back-door criterion was used to find a valid estimand, and the causal effect was estimated using linear regression with confounders as conditioning variables. The resulting ATE was calculated thus the ATE is -0.0363, showing a risk reduction of 3.63% in heart disease because of physical activity. The Bootstrap confidence intervals [0.039,0.0337] proved the consistency of this result.

5.3 Propensity Score Matching (PSM)

Using Propensity Score Matching (PSM), we estimated how physical activity impacts heart disease using comparable treated and untreated groups based on their propensity scores. Propensity scores, which portray the probability of receiving the treatment given confounders (X, which includes age, BMI, smoking, and physical health), were precalculated for all persons. Nearest neighbor matching was selected for identifying matched treated individuals with the same or nearly the same propensity scores as the matched controls, to separate the selection bias. The Average Treatment Effect (ATE) was then calculated as:

$$\text{ATE} = \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0],$$

where $e(X)$ and $e(Y)$) are in turn employed to signify the respective levels of achievements for the pairs of matched cases and controls. Using this method, we have obtained an ATE of -0.0369, which implies a 3.69% defense against the risk of heart disease for people who are physically active. Matching quality was checked by looking at the overlap in pairwise propensity scores and the comparability of the covariates across the treated and matched control groups.

5.4 Inverse Probability of Treatment Weighting (IPTW)

To estimate the causal effect of physical activity on heart disease using the Inverse Probability of Treatment Weighting (IPTW), we created a weighted pseudo-population where the covariates were balanced between treated (physically active) and untreated (physically inactive) groups. The treatment assignment was independently controlled for confounders like age, BMI, smoking, and physical health. We computed propensity scores that served as weights to adjust for the confounding factors as mentioned earlier.

A weighted least squares regression analysis with heart disease as the outcome and physical activity as the predictor was conducted. The weights that were gotten from the propensity score allowed us to control for confounders and isolating the effect of physical activity. The IPTW analysis resulted in an Average Treatment Effect (ATE) of -0.0349. This showed a 3.49% decline in heart disease risk for physically active individuals. Through balance diagnostics, we came to the conclusion that the covariates were well-balanced after the weighting was done. This is a testament to the robustness and reliability of the causal effect estimation.

5.5 S-Learner with Random Forest

The S-learner framework was integrated with Random Forests to estimate Heterogeneous Treatment Effects (HTE). In this approach, the treatment variable (*PhysicalActivity*) was added as a feature to the input dataset, enabling the model to learn both treated and control outcomes. Random Forests were chosen as the base learner due to their capability to handle nonlinearities, high-dimensional data, and interaction features effectively.

The model was trained on observational data, with the treatment variable and other covariates as predictors and *HeartDisease* as the outcome. Predictions for treated (A=1) and control (A=0) groups were obtained by altering the treatment variable while keeping other features constant. The individual treatment effect was computed as the difference between the two predictions for each subject.

5.6 ERMSE for Robustness Evaluation

The S-learner model's robustness was evaluated using ERMSE (Estimated Root Mean Squared Error), as suggested by Shalit et al. (2017). ERMSE measures the consistency of the model's predictions across treatment and control groups by comparing observed and predicted outcomes. It serves as a reliable metric to assess how well the model generalizes and estimates treatment effects.

ERMSE was computed separately for treated and control groups, representing prediction error for each group. The averaged ERMSE was used as a general measure of the model's robustness in estimating HTE within an observational setting.

5.7 SHAP Analysis for Feature Interpretability

To explore the dimensions of treatment heterogeneity, SHAP (SHapley Additive exPlanations) analysis was performed. SHAP values quantify the marginal contribution of each feature to the predicted HTE by assessing the effect of removing that feature in multivariable interactions. This analysis helps identify the features driving treatment heterogeneity.

SHAP analysis identified *PhysicalHealth*, *BMI*, and *Smoking* as key features contributing significantly to treatment effect variation. Visualizing feature contributions through SHAP analysis provided contextually relevant interpretations of the model's predictions, essential for health applications. These insights can inform targeted interventions by highlighting the sources of heterogeneity.

6 Data description and cleaning

The dataset used for this project is downloaded from Kaggle and claims to be from a CDC survey. Health-related variables include physical activity, lifestyle habit data, and heart disease status. Categorical and continuous variables add up to more than 310,000 observations.

6.1 Data Cleaning

- **Missing Values**: The dataset was checked for missing data, and the output confirmed no missing values. This step is crucial as it assures the integrity of the dataset for further analysis.
- **Outliers**: For continuous variables such as *BMI*, *PhysicalHealth*, and *SleepTime*, boxplots were used to detect outliers. Extreme values were identified but not removed or modified in the present notebook. However, the existence of outliers was considered in the interpretation of the results.

6.2 Data Transformations

- **Categorical Variables**: Variables such as *PhysicalActivity*, *Smoking*, and *GenHealth* were converted into dummy or binary variables to enable feasible statistical analysis.
- **BMI**: BMI (Body Mass Index) and poor physical health were utilised to place individuals within the context of their weight (BMI classification and their health in turn). Being overweight or not was assigned value 1 hence poor physical health or otherwise was assigned value 0.
- **One-Hot Encoding**: Categorical variables with more than two categories, such as *AgeCategory* and *GenHealth*, were one-hot encoded to facilitate analysis.
- **HealthRiskIndex**: A composite index created by combining several health markers (e.g., BMI, PhysicalHealth). The distribution of this index was plotted to examine its dispersion in the dataset, along with measures of central tendency and variability. Relevant frequency distributions were also plotted to assess the general health conditions of the population.

6.3 General Insights

- **Prevalence of Heart Disease**: A clear pattern indicates that physically active individuals have a significantly lower prevalence of heart disease compared to physically inactive individuals.
- **Age and Physical Activity**: Younger individuals are more likely to engage in regular physical activity, while such activities tend to decline with age, likely due to physical impairments that limit exercise capabilities.
- **BMI Distribution**: The BMI distribution reveals that a large portion of the population falls within the overweight or obese categories, aligning with global trends indicating a rise in obesity.

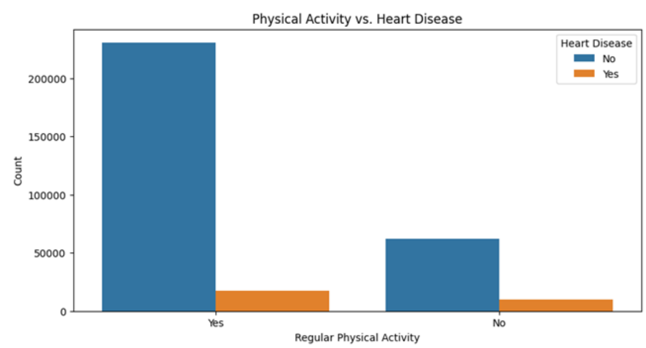

Figure 1

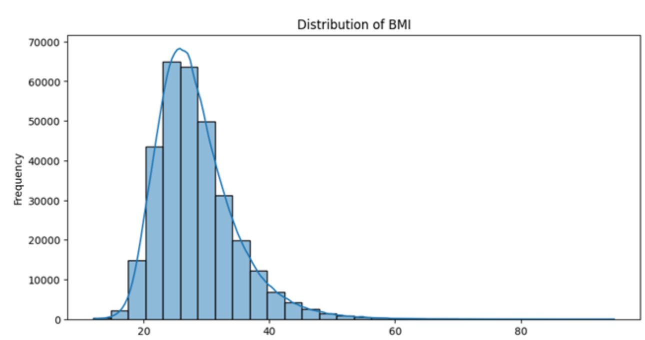

Figure 2

6.4 Preprocessing for HTE

Some of the preprocessing activities carried o
dataset for HTE analysis again to ensure prepared
patibility with models and data quality. Certain re
which were created as part of feature engineeri
Essential demographic, health-related, and behav
as *BMI*, *PhysicalHealth*, *Smoking*, and derived *Ge*
were retained for analysis. The target variable, *I*
converted to binary format (1 for "Yes" and 0 fc
classification.

Categorical variables such as *AgeCategory*, *R*
betic were numerically encoded to make them eli
learning algorithms. Standardizing the data struct
evant features and prepared it for robust treatmen
These preprocessing steps were essential to deri
sights from the subsequent HTE analysis.

7 Experimental Setup and Results

Experimental Setup

Objective: To estimate the causal effect of ph
heart disease risk while addressing confounding
Exploratory Data Analysis (EDA): EDA provic
into the dataset, highlighting relationships and po
Key observations included:

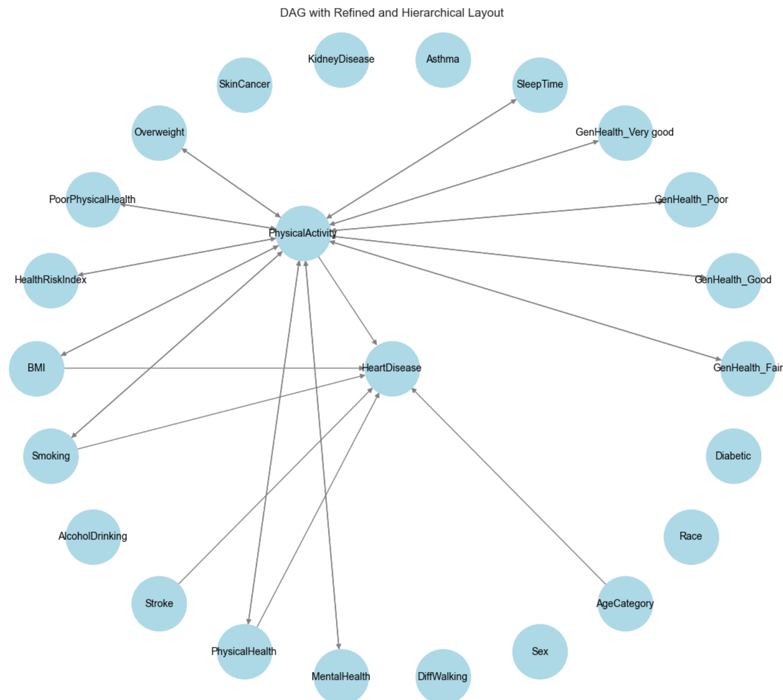

Figure 3: Directed Acyclic Graph.

7.1 Methods its Results

DAG Construction and Confounder Identification: The analysis controlled for confounders identified through EDA and the DAG, such as *BMI*, *Smoking*, and *PhysicalHealth*. Three causal inference methods were implemented:
Role of the DAG: The DAG (Figure 3) visually represented the causal relationships between variables in the dataset. It helped identify confounders influencing both the treatment (*PhysicalActivity*) and the outcome (*HeartDisease*), ensuring valid causal estimates.
Figure 3: Directed Acyclic Graph (DAG) The DAG (Figure 3) illustrates the causal relationships between treatment (*PhysicalActivity*), outcome (*HeartDisease*), and confounders (*BMI*, *Smoking*, *PhysicalHealth*).

Baseline Analysis: The bare comparison at baseline considered the frequency of heart disease between active (A=1) and inactive (A=0) persons without confounding considerations. This unadjusted comparison suggested that inactive individuals have a higher prevalence of heart disease that implies physical activity as a protective factor. These results are considered probably biased because of the confounding variables such as age, BMI, and smoking. Therefore, the conclusions may not be valid anymore. Baseline has served to indicate the above as a reference point to stress the importance of stronger causal inference techniques.

Balancing Covariates: Logistic regression was used to compute propensity scores for estimating the likelihood of being physically active among individuals based on confounding factors such as smoking, Body Mass Index (BMI) and physical health. These derived scores were used to equate treated (A=1) and control (A=0) groups. Matched pairs of individuals with similar propensity scores help to create comparability, while balanced weighting takes care of possible differences in confounder distributions in the whole population.

Estimating Causal Effects: Advanced causal inference methods were applied to estimate the causal effect of physical activity on heart disease. A weighted least squares regression model used inverse probability of treatment weighting (IPTW) to simulate a pseudo-population where confounders were balanced. Additionally, regression adjustment directly modeled the relationship between physical activity and heart disease while conditioning on identified confounders. These methods provided consistent ATE estimates across approaches.

Validation and Robustness: Checked and validated estimates assuring diagnostic purity. Standard mean differences (SMD) were used in judging covariate balance prior to as well as after adjustment, with final SMD values being reduced below 0.01, approving effective balance. Bootstrap confidence intervals quantified the uncertainty in the estimates, further strengthening them.

Method	ATE Estimate	95% Confidence Interval
IPTW	-0.0349	[-0.037, -0.033]
PSM	-0.0369	-
Backdoor Adjustment	-0.0363	[-0.039,-0.0337]
Subset Refutation Test	-0.0363	-
Bootstrap	-0.0363	[-0.0390, -0.0337]

Table 1: ATE estimates for the effect of physical activity on heart disease.

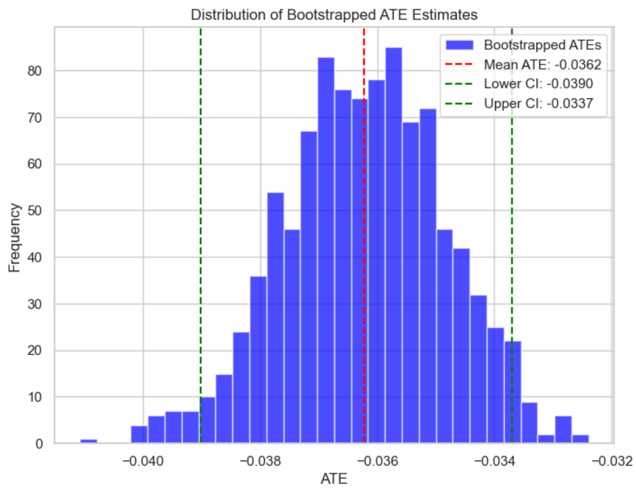

Figure 4: Distribution of Bootstrapped ATE Estimates

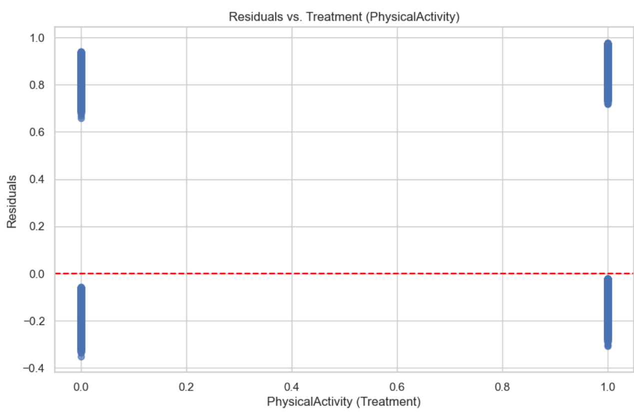

Figure 5: Residuals vs. Treatment validating model fit

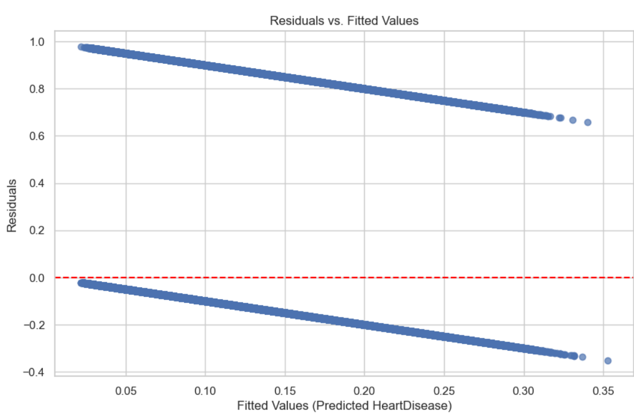

Figure 6: Residual plot validating regression model assumptions

Traditional Subgroup Analysis with PSM: At first, Propensity Score Matching (PSM) was applied to investigate the variation in treatment effects across the pre-defined subgroups based on confounding important factors defined in the DAG such as BMI, Smoking, PhysicalHealth. This was by matching individuals within these subgroups and giving us subgroup specific estimates of HTE, which offer an initial baseline into the treatment variation.

Random Forests: The S-Learner Approach S-learner with random forests has been applied to assess individual-level heterogeneous treatment effects (HTE) rather than subnetworks from subgroup analysis. The treatment variable was included in the dataset that contains variables, as an example *PhysicalActivity*, and allowed the model to learn the treated (1) and the control (0) outcomes for one another. Treatment was changed in the prediction by all other features, understanding that the model would apply a different prediction for the treated and control scenarios. The Average Treatment Effect (ATE) derived by the mean of *HTE = -0.0021* , whereas the *max HTE = 0.179*, the *min = -0.18*, and the *standard deviation = 0.018*. This suggested that physical activity, on average, might not have a big impact with respect to heart disease based on the dataset we worked on. Yet, the HTE density suggests a more significant individual variation with respect to the ATE due to the fact that they are concentrated close to zero, as seen in Figure 4.

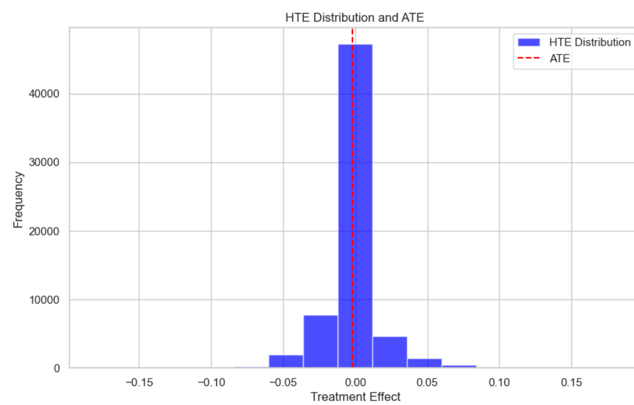

Figure 7: HTE Distribution and ATE through S-learner

SHAP Analysis of the Feature: SHAP analysis was performed for interpretation of the drivers with respect to heterogeneity treatment estimation. In summary, SHAP values measure the marginal contributions of individual features to variability in heterogeneous treatment effects (HTE). The analysis did point out a few key predictors: *PhysicalHealth*, *BMI*, *Smoking*, and *AgeCategory*, which are among the most significant contributors to treatment heterogeneity. For such a model, this summary would provide an interpretative visualization in the form of a SHAP summary plot that most effectively demonstrates importance as seen in Figure 5.

7.2 Evaluation Metrics

ATE Metrics: Back-Door Rehabilitation brought 0.0363 to ATE, respectively, and PSM and IPTW also reached the same negative values of 0.0369 and 0.0349. Addressing techniques such as the bootstrap confidence intervals and the fairness diagnostics treated

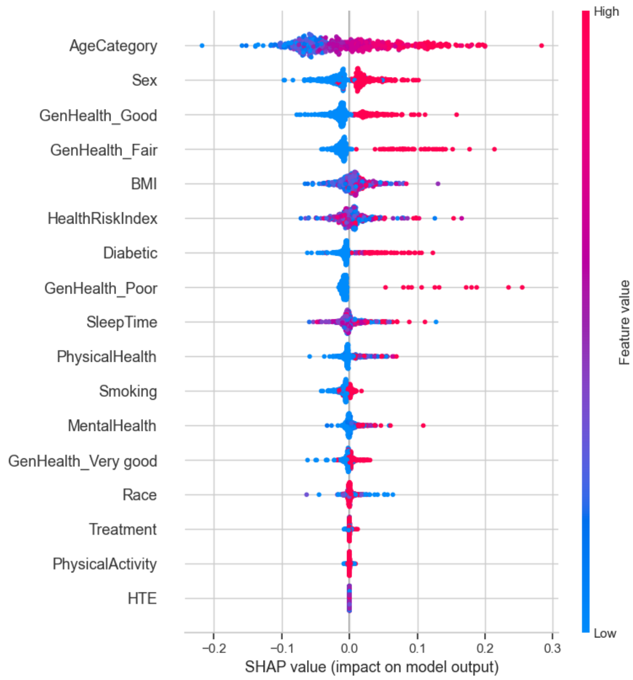

Figure 8

the estimation quite successfully. These were the strategies that helped to identify that physical activity may reduce the chance by about 3.5%, informing the necessary evidence for public health interventions.

The Approximated Root Mean Square Error (ERMSE)

ERMSE was applied to evaluate the S-learner performance while Random Forests predicted results in treated and control groups' outcome variables. It was derived from both the treated and control populations wherein their observed and predicted values of outcomes were compared. Finally, the mean of both values provided an overall view of the ERMSE concerning the robustness of the model.

- **Treated Group ERMSE:** 0.2621
- **Control Group ERMSE:** 0.3407
- **Overall ERMSE:** 0.3014

These results showed how reliable the model could predict similar outcomes among the treated and control groups and thus validate the confirmed robustness of the S-learner framework.

SHAP Assessment:

The SHAP plot highlights the most prominent drivers of heterogeneity in treatment. While *AgeCategory* was the most pervasive feature, *GenHealth*, *BMI*, *PhysicalHealth*, and *Smoking* were next in line as major determining factors for HTE estimates. These characteristics shape significant portions of HTE estimates as evidenced by their individual SHAP values in the analysis of their impact on treatment effects at the individual level.

Conclusions

(1) **Physical Activity Reduces Heart Disease Risk:** The causal effect of physical activity on heart disease was consistently significant across all methods, demonstrating its protective role.
(2) **Importance of Controlling for Confounders:** The baseline comparison overstated the effect of physical activity due to confounding. Adjusting for *BMI*, *Smoking*, and *PhysicalHealth* revealed a more accurate estimate.
(3) **Targeted Interventions:** HTE analysis highlights opportunities for targeted interventions, particularly for smokers and younger individuals, to maximize the benefits of physical activity.

8 Discussion and ideas for next steps:

References

(1) Pytlak, K. (2021). Personal Key Indicators of Heart Disease [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
(2) Künzel, S. R., Sekhon, J. S., Bickel, P. J., Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116
(3) Ling, Y., Upadhyaya, P., Chen, L., Jiang, X., Kim, Y. (2022). Emulate Randomized Clinical Trials using Heterogeneous Treatment Effect Estimation for Personalized Treatments: Methodology Review and Benchmark. Journal of Biomedical Informatics, 104256.
(4) Gąsior, J. S., Młyńczak, M., Williams, C. A., Popłonyk, A., Kowalska, D., Giezek, P., & Werner, B. (2023). The discovery of a data-driven causal diagram of sport participation in children and adolescents with heart disease: a pilot study. *Frontiers in Cardiovascular Medicine, 10.* https://doi.org/10.3389/fcvm.2023.1247122
(5) Myers, N. D., Brincks, A. M., & Lee, S. (2024). Direct and indirect causal effects of an individual Randomized Physical Activity-Promoting intervention: a Substantive-Methodological synergy. *Measurement in Physical Education and Exercise Science*, 1–19. https://doi.org/10.1080/1091367x.2024.2392563
(6) Shi, J., & Norgeot, B. (2022). Learning Causal Effects From Observational Data in Healthcare: A Review and Summary. *Frontiers in Medicine, 9:864882.* https://doi.org/10.3389/fmed.2022.864882
(7) Qiu, P., Wu, J., Li, M., Zhao, Z., & Wang, Q. (2024). Causal inference between physical activity and chronic diseases: insights from a two-sample Mendelian randomization study. *Sports Medicine and Health Science.* https://doi.org/10.1016/j.smhs.2024.09.002
(8) Wade, K. H., Richmond, R. C., & Davey Smith, G. (2018). Physical activity and chronic disease: Unpacking causal relationships. *British Journal of Sports Medicine, 52*, 891–892.
(9) Pan, W., Bai, H. (2015). Propensity score interval matching: Using bootstrap confidence intervals for accommodating estimation errors of propensity scores. *BMC Medical Research Methodology, 15(1), 53.* https://link.springer.com/article/10.1186/s12874-015-0049-3
(10) Austin, P. C., Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting

(IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34(28), 3661–3679*. https://link.springer.com/article/10.1007/s00392-021-01846-7