Missouri University of Science and Technology

Department of Mathematics and Statistics

Assessing Causal Relationships Between Cardiovascular Risk Factors and Heart Disease Using Bias-Adjusted and Sensitivity-Robust Methods Using Cleveland Heart diseases Data

By:

Dejene Chala, Irene Kidd, Jacob Meyers, Quinlin Neuhaus

Submitted to Dr. Robert Paige

December 5, 2025

Summary

Coronary heart disease (CHD) remains the leading cause of morbidity and mortality worldwide (Lu & Lan, 2022). While many observational studies have documented associations between risk factors such as smoking, hypertension, dyslipidemia, diabetes, and obesity with coronary heart disease (CHD), association is not causation. In this project, we applied causal inference approaches using the Cleveland UCI heart-disease dataset to estimate the causal effect of modifiable risk factors on CHD. We have employed causal DAGs, logistic regression, bootstrapping, propensity score methods, inverse probability weighting, doubly robust estimators, targeted maximum likelihood estimation (TMLE) and mediation analyses. The objective outcome variable was the presence of heart disease, which was treated as a binary dummy variable. Fasting blood sugar was identified as the exposure, while age and sex were identified as confounders, and cholesterol was identified as a mediator based on a review of recent studies. Exploratory data analyses (correlation matrix and correlation heatmap) were done for a quick check of the nature of the relationship between variables, which found some strong correlation between the outcome, confounders, and some variables that were regarded as symptoms of the presence of the outcome variable. The correlation between exposure and outcome variable was non-significant, which might be attributed to the effect of confounders or mediator variables that need to be adjusted for. So, we explored the causal inference method that helps to adjust for the effect these variables so that we can extract the causal effect of treatment/exposure on the outcome variable of interest. Key outputs include causal effect estimates, directed acyclic graphs, and sensitivity analyses to evaluate robustness. The results of logistic regression of outcome on potential exposure variable fasting blood sugar has an estimated log-odds ratio of 0.88 with p-value 0.723. Bayesian bootstrapping with weak priors was done to estimate the ATE, the posterior mean of ATE was -0.0406, with 95% CI of (-0.1657, 0.0786). Propensity score methods yield causal ORs of 1.07122 with p-value 0.674 and 0.91960 with p-value 0.806 for the exposure model and outcome model respectively. TMLE was performed using the SuperLearner and tmle packages where the estimated relative risk of treatment vs control was 1.023, and the causal OR was 1.044, with p-value of 0.906. The results of parametric estimation for the mediation analysis estimates that the NDE is -0.0746, with 95% CI (-0.212, 0.075), NIE is -0.0031with 95% CI (-0.517, 0.031), TE is -0.0777 with 95% CI (-0.220, 0.067). Overall, no significant causal effect was found on the relationship between fasting blood sugar and presence of coronary heart disease in the Cleveland UCI heart disease dataset.

Acknowledgments

We would like to express our deepest gratitude to **Dr. Robert Paige** for giving us the invaluable opportunity to work on this project and for his continuous support and guidance from the development of the proposal to the completion of the project. His insightful suggestions on appropriate methodologies, constructive feedback on our comments and questions during lectures, and his willingness to respond to our inquiries outside class hours have been instrumental to our learning and the success of this work.

Table of contents

Summary ii

ACKNOWLEDGMENTS III

TABLE OF CONTENTS IV

1. INTRODUCTION 2

1.1. Background of the Study 2

1.2. Objectives of the Study 2

2. METHODS 3

2.1. Meta Data and Causal DAG 3

2.2. Do-Calculus Using Bayesian-Network 4

2.3. Logistic Regression 5
2.3.1. Conceptual Framework and Assumptions 5
2.3.2. Statistical Modeling 5
2.3.3. Assessment of Model Assumptions and Diagnostics 6

2.4. Bootstrap-A Bayesian Approach 7
2.4.1. The Model 7
2.4.2. The Bootstrap Framework 8
2.4.3. Computation of the Average Treatment Effect 8

2.5. Propensity score methods 9
2.5.1. Theory 9
2.5.2. Computation 9

2.6. ATE, G-computation, AIPTW, and TMLE 10
2.6.1. Computational Procedures 10

2.7. Mediation Analysis 12
2.7.1. Motivations and Reasoning 12
2.7.2. Model Specifications 12

3. RESULTS 13

3.1. Causal DAG and data exploration 13

3.2. Results of Do-Calculus with Netica 15

3.3. Results of Logistic regression 15

3.4. Bootstrap estimation procedure 17

3.5. Propensity Score Methods Results 18

3.6. Targeted Maximum Likelihood Estimation, G-computation and AIPTW 19

3.7. Mediation Analysis Results 20
3.7.1. Parametric Linear Model 20

4. CONCLUSION 21

References 23

APPENDIX: R CODE USED IN THE STUDY 24

CAUSAL EFFECT OF FASTING BLOOD SUGAR ON HEART ATTACK 24

Introduction 24

Data Import 24

Variable Definitions and Data Preparation 24

Descriptive Statistics 25

Model 1: Total Effect of FBS (Adjusted for Age and Sex) 27

Model 2: Direct Effect of FBS (Adjusted for Age, Sex, Cholesterol, and Resting BP) 28

Comparison of FBS Effect Across Models 30

Model Diagnostics 31

Interpretation 32

MEDIATION USING PACKAGE 33

1. Introduction

1.1. Background of the Study

Cardiovascular diseases, particularly ischemic heart disease, remain a major cause of global morbidity and mortality. According to the Global Burden of Disease report, ischemic heart disease has been the leading cause of death and disability globally for decades (Mensah et al., 2023) . Studies indicated that modifiable risk factors, including smoking, hypertension, dyslipidemia, diabetes, obesity, poor diet, and psychosocial stress, account for over 90% of the population-attributable risk for myocardial infarction (Yusuf et al., 2004). Other cohort studies, such as (Seeman et al., 1993) and (Tervahauta et al., 1995), have further shown that biological and psychosocial factors significantly contribute to cardiovascular risk. All these studies establish associations, not causation.

Association studies can be affected by confounding, selection bias, and measurement error. Modern causal inference methods, as formalized in Pearl's structural causal model (SCM) framework (Austin, 2011) and epidemiological approaches discussed by Hernán and Robins (Hernán & Robins, 2018), enable researchers to formally express causal assumptions, identify valid adjustment sets, and estimate causal effects under explicit assumptions. This study will therefore use the Cleveland UCI heart-disease dataset to estimate the causal relationships between key risk factors and CHD using robust causal inference methods.

1.2. Objectives of the Study

The primary objective of this study is to estimate the causal effects of selected modifiable risk factors on CHD.

Specific Objectives

1. Explore different parametric and non-parametric methods of estimating the causal effect of the exposure on the outcome adjusted for appropriate confounders.
2. Evaluate uncertainty using bootstrapping and influence-function-based inference.
3. Obtain bias adjusted treatment effect estimates using propensity score, mediation, G-Computation, AIPTW, and TMLE.

2. Methods

We have used different graphical and numerical methods in order to achieve the intended objectives.

2.1. Meta Data and Causal DAG

We have used Cleveland UCI heart diseases data that has 303 cases with 14 variables collected.
The list of variables and their descriptions was given in table 1.

List of all variables in the data with their descriptions were given in table 1.

Table 1: Meta data

Variable	Description
Age	Age in years.
Sex	1 = male, 0 = female.
Cp	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal, 4 = asymptomatic).
Trestbps	Resting blood pressure in mm Hg.
Chol	Serum cholesterol in mg/dl.
Fbs	Fasting blood sugar >120 mg/dl (1 = true, 0 = false).
Restecg	Resting electrocardiographic results (0 = normal, 1 = ST–T abnormality, 2 = LV hypertrophy).
Thalach	Maximum heart rate achieved (bpm).
Exang	Exercise-induced angina (1 = yes, 0 = no).
Oldpeak	ST depression induced by exercise relative to rest.
Slope	Slope of peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping).
Ca	Number of major vessels (0–3) colored by fluoroscopy.
Thal	Thallium stress test result (3 = normal, 6 = fixed defect, 7 = reversible defect).
Target	Heart disease presence (0 = no disease, 1 = disease).

The outcome variable of interest in this study is the target(which indicated the presence or absence of Coronary Artery Disease, CAD). The variables considered listed above include confounders, potential risk factors, and manifestations/symptoms of the presence of CAD. The following is a detailed discussion of the nature of the relationships present.

The variables age, sex, resting blood pressure, cholesterol level, and fasting blood sugar are widely recognized predictors of coronary artery disease (CAD). Advancing age and male sex significantly elevate CAD risk, as shown in large epidemiologic studies (Rodgers et al., 2019; Kim et al., 2023). Here age and sex can be regarded as measured confounders since we cannot control them, and the variables have no parents in the DAG. Hypertension (high blood pressure) is a major modifiable determinant of CAD(Unger et al., 2020).

Clinically, CAD manifests through chest pain (angina), reduced exercise tolerance, abnormal maximum heart rate response, and exercise-induced angina (Gulati et al., 2021). So, Chest can be classified as a symptom of heart disease. Stress testing frequently reveals ischemia through ST-segment depression or abnormal ST-segment slope, which correlate strongly with obstructive coronary disease (Vilcant & Zeltser, 2023). The severity of CAD is reflected in the number of major coronary vessels involved, typically assessed through coronary angiography. Resting electrocardiographic abnormalities may indicate prior myocardial injury or active ischemia, contributing to diagnostic evaluation and risk stratification (Shahjehan & El-Sherief, 2024). The DAG is then constructed based on the above nature of relationships between the variables.

2.2. Do-Calculus Using Bayesian-Network

Bayesian Network (BN) was constructed in **Netica** to encode the assumed causal relationships among relevant clinical variables based on the causal DAG established. The following steps were followed to conduct the do-calculus. The node for each of the relevant variables were created, states were created for categorical variables and continuous variables were discretized.

The modelling steps:

1. The confounder variables **age** and **sex** were designed to have directed edges as *age → fbs*, *sex → fbs*, *age → target*, and *sex → target*. This structure ensures that Netica appropriately adjusts for these backdoor paths when estimating the causal effect of fbs.

2. **Chol** (serum cholesterol) was modeled as a **mediator** of the fbs–target relationship. Clinical evidence suggests that abnormal glucose regulation contributes to dyslipidemia, which subsequently impacts cardiovascular risk. This mechanism was represented by the directed pathway *fbs → chol → target*, while also retaining a direct arrow *fbs → target* so that both direct and indirect causal effects could be quantified.

3. **Trestbps** (resting blood pressure) was incorporated as **effect modifier**. Rather than confounding the fbs–target relationship, these variables alter its magnitude. In a BN framework, effect modification is implemented by allowing the conditional distribution of the outcome to depend jointly on the exposure and the modifier.

Therefore, *trestbps* was set as additional parent nodes of *target*, enabling Netica to estimate heterogeneous causal effects across levels of these variables.

After the structure was defined, the cleaned dataset was imported as a case file and Netica's **Learn CPTs** procedure was used to estimate model parameters. Continuous variables were discretized into clinically meaningful intervals to enhance interpretability and computational stability. The network was then compiled to enable inference.

Causal effects were estimated using Netica's **intervention (do-operator) functionality**. An intervention on fbs, denoted *do(fbs = x)*, forces the network to simulate an active manipulation of fbs while severing all incoming edges to fbs, thereby eliminating confounding by age and sex. The resulting distribution *P(target | do(fbs = x))* represents the estimated causal effect of modifying fbs. Effect modification was assessed by comparing interventional results across strata of trestbps and ca, and mediation was examined by evaluating how the distribution of chol responds under intervention. This BN-based methodology provides a coherent and transparent framework for causal inference that explicitly incorporates confounding control, mediation pathways, and effect heterogeneity.

2.3. Logistic Regression

2.3.1. Conceptual Framework and Assumptions

The aim is to find the causal effect of blood sugar (FBS) on heart attack occurrence using logistic regression, while adjusting for confounders and accounting for potential mediators. The exposure of interest is **fbs** with binary outcome **target,** Confounders: **age, sex**. Mediators: **chol** (serum cholesterol), **trestbps** (resting blood pressure). For valid causal interpretation, the following assumptions are invoked: (i) no unmeasured confounding of the FBS–heart attack relationship after conditioning on age and sex; (ii) no unmeasured confounding of the FBS–mediator and mediator–outcome relationships; (iii) correct model specification; and (iv) no substantial measurement error in the variables.

2.3.2. Statistical Modeling

The modeling strategy is structured to: Estimate the confounder-adjusted total effect of fasting blood sugar on heart attack and explore how adjustment for potential mediators (cholesterol and resting blood pressure) reduces the FBS–heart attack association.

A. Confounder-Adjusted Total Effect Model (Model 1)

To estimate the total effect of exposure on Y, we fit a logistic regression model specified as:

$$logit[P(target\ =\ 1\mid fbs, age, sex)]\ =\ \beta_0\ +\beta_1 * fbs\ +\ \beta_2 * age\ +\ \beta_3 * sex \qquad (1)$$

Where $logit(p) = \log\left(\frac{p}{1-p}\right)$. In this model, β_1 represents the log-odds ratio for heart attack comparing individuals with (fbs = 1) to those without (fbs = 0), adjusted for age and sex. The corresponding odds ratio for the total effect is exp (β_1).

B. Mediation Model Including Cholesterol and Blood Pressure (Model 2)

To assess the role of cholesterol and resting blood pressure as mediators, we extend the logistic regression model to include chol and trestbps as additional covariates:

$$logit\big[p\big(y = 1\big|(A, C, M)\big)\big] = \gamma_0 + \gamma_1 fbs + \gamma_2 age + \gamma_3 sex + \gamma_4 chol +\ \gamma_5 testbps \qquad (2)$$

Where $where\ (A, C, M) = (fbs, age, sex, chol, trestbps)$

In this expanded model, γ_1 estimates the log-odds ratio for the direct effect of FBS on heart attack, conditional on age, sex, cholesterol, and resting blood pressure. Under the causal assumptions, exp (γ_1) can be interpreted as an approximation of the controlled direct effect of FBS on heart attack, controlling for the mediators.

Comparing the FBS coefficient between Model 1 (β_1) and Model 2 (γ_1) allows us to evaluate the degree of weakening in the association after accounting for the mediators. A substantial reduction in the magnitude of the FBS odds ratio when mediators are included would be consistent with some of the total effect being transmitted through cholesterol and/or blood pressure.

C. Model Fitting and Estimation

Models will be estimated using maximum likelihood estimation. The primary parameter of interest is the odds ratio for FBS:

- In Model 1: OR_{total} = exp (β_1) , interpreted as the odds ratio for heart attack for individuals with elevated FBS vs. normal FBS, adjusted only for confounders.
- In Model 2: $OR_{direct} \approx$ exp (γ_1), interpreted as the odds ratio for heart attack associated with elevated FBS after controlling for both confounders and mediators.

2.3.3. Assessment of Model Assumptions and Diagnostics

Several diagnostics will be performed to assess model adequacy and assumptions:

- **Multicollinearity**: Variance inflation factors (VIFs) or correlation matrices will be examined to identify potential multicollinearity between predictors, especially between chol and trestbps or between age and the mediators.
- **Goodness-of-fit**: Overall model fit will be evaluated using measures such as the Hosmer–Lemeshow test and pseudo R-squared statistics. ROC curve was used to evaluate the predictive performance of the model.

2.4. Bootstrap-A Bayesian Approach

The primary causal question is: what is the effect of (fbs) on the probability of having a heart attack, adjusting for confounders, and considering potential mediators, after assuming weak priors on the parameters? The variable definitions are as before, and the nature of relationship declared in logistic regression were also adopted here. The core idea is to build a Bayesian logistic regression in Stan, apply nonparametric bootstrapping at the data level, and for each bootstrap sample fit the same Stan model. Causal effects (e.g., average treatment effect of fbs on the probability of heart attack) are then computed from the posterior draws for each bootstrap replicate. Finally, ShinyStan is used to explore and diagnose each Stan fit.

Assumptions for valid inference. **C**onsistency, Positivity, Conditional exchangeability are assumed.

2.4.1. The Model

The core statistical model for the outcome is a Bayesian logistic regression linking the log-odds of heart attack to fbs, age, sex, chol, and trestbps. For subject i = 1, …, N, the model can be written as:

$$logit\big(P(Y_i = 1)\big) = \alpha + \beta_{fbs} * fbs_i + \beta_{age} age_i + \beta_{sex} * sex_i + \beta_{chol} * chol_i + \beta_{trestbps} * trestbps_i$$

with Y_i modeled as a Bernoulli random variable with probability given by the inverse-logit of the linear predictor. Weakly informative (Normal) priors are placed on the intercept and regression coefficients. The structure in stan contains**: Data**: number of observations N; binary outcome y; binary exposure fbs; continuous age, chol, trestbps; binary sex. **Parameters**: $\alpha, \beta_{fbs}, \beta_{age}, \beta_{sex}, \beta_{chol}, \ \beta_{trestbps}.$ **Model block**: specification of the logistic regression and priors. This model provides posterior draws for all regression coefficients, from which we can compute causal contrasts on different scales (risk difference, risk ratio, or odds ratio).

2.4.2. The Bootstrap Framework

The bootstrap layer is applied outside Stan at the data level. The steps are:

1. Start with the original dataset containing Y, exposure, confounders and mediators.
2. For each bootstrap replicate b = 1, …, B:
   - Draw a bootstrap sample by sampling N rows from the original dataset with replacement.
   - Construct the list of data inputs required by the Stan model (y, fbs, age, sex, chol, trestbps, and N).
   - Fit the Stan model to this bootstrap dataset using rstan, obtaining posterior draws for all parameters.
   - From the posterior draws, compute a summary measure of the causal effect of interest (for example, the average treatment effect of fbs on the probability of heart attack).
3. Aggregate the estimates across bootstrap replicates: store one causal effect estimate per bootstrap sample (or store full posterior summaries for each replicate), and then compute bootstrap means, standard deviations, and confidence or credible intervals.

2.4.3. Computation of the Average Treatment Effect

A natural estimand in this setting is the average treatment effect (ATE) on the risk scale, defined as the average difference in the probability of heart attack if all individuals were set to fbs = 1 versus if all individuals were set to fbs = 0, averaging over the empirical distribution of covariates in the dataset. Using posterior draws from Stan:

For each posterior draw of $\alpha, \beta_{fbs}, \beta_{age}, \beta_{sex}, \beta_{chol}, \beta_{trestbps}$), compute two sets of predicted probabilities for each individual.

- $P(1)_i$: predicted probability of $Y_i = 1$ if fbs_i were set to 1, holding age, sex, chol, and trestbps at their observed values.
- $P(0)_i$: predicted probability of $Y_i = 1$ if fbs_i were set to 0, again holding other variables at their observed values.

1. For each draw, compute the individual-level differences $P(1)_i - P(0)_i$: and average them over all individuals to obtain a draw-specific ATE on the probability scale.
2. Repeat over all posterior draws and summarize the distribution (posterior mean, median, and credible interval).

Within each bootstrap sample, this yields a posterior distribution for the ATE. The mean or median of this posterior can be taken as the ATE estimate for that particular bootstrap replicate. Across B bootstrap samples, the B ATE estimates provide a bootstrap distribution that reflects both sampling variability (through resampling) and model-based uncertainty (through the posterior).

2.5. Propensity score methods

2.5.1. Theory

Most causal data science methods adjust a treatment effect with the set of sufficient confounders, however in the case of many confounders, it may be helpful to simplify these into one single confounder. In fact, if a set of sufficient confounders exists, then the univariate propensity score must also be a sufficient confounder. While this property holds exactly for a known function of the confounders, it will approximately hold for estimated parameters of the confounders, most commonly done with logistic regression. These propensity scores estimate an individual's probability of receiving the treatment based on the set of sufficient confounders. The propensity scores can then used as an ordinary confounder, to match treated and untreated individuals, or to standardize in the exposure model.

2.5.2. Computation

As with all other methods, we assume the following variable notations: Treatment: **A = fbs**; Outcome: **Y = target**; Confounders: **W = (age, sex).**

Using a logistic regression of $logit\big(P(A = 1)\big) = \beta_0 + \beta_1 age + \beta_2 sex$ to obtain fitted values for the propensity scores, $e(W)$. To check the positivity assumption, look at the densities of the propensity scores for the treated and untreated groups. A high density near 0 or 1 that is unaccompanied by the other group could be a positivity violation.

A. Standardization with exposure modeling

Using the propensity score, e(W), the observations can be weighted. This makes the adjustment that including the confounders would have had, weighting observations to equate the distributions of confounders among the treated and untreated groups, using only the propensity score.

B. Propensity scores in the outcome model

In the outcome model, the propensity scores themselves are used in a model in direct substitution of the set of sufficient confounders, along with the treatment. Because of our data's binary target variable, the outcome model used was a logistic regression of $logit\big(P(Y = 1)\big) = \beta_0 + \beta_1 fbs + \beta_2 e(W)$.

C. Matching on propensity score

One final application of the propensity score method is matching. In traditional matching, a researcher might attempt to match each treated individual to one or more untreated individuals with the same confounders. If the number of confounders is large, this may not be possible, and we match on propensity score instead of the entire set of confounders. Matching provides an intuitable strategy to compare treated and untreated groups and can be used with propensity score.

2.6. ATE, G-computation, AIPTW, and TMLE

2.6.1. Computational Procedures

Variable notations: **Treatment: A = fbs** (1 = high fasting blood sugar, 0 = normal); **Outcome: Y = target** (1 = heart disease present, 0 = none); **Confounders: W = (age, sex).**

1. Average Treatment Effect (ATE)

Define potential outcomes: **Y1**: outcome if everyone had **fbs = 1, Y0**: outcome if everyone had **fbs = 0.** Then the ATE is: ATE $= E[Y_1 - Y_0]$

Under identification assumptions (consistency, exchangeability given W, positivity), this equals: ATE $= E_W\big[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]\big]$. In our dataset: $E[Y \mid A = a, W]$ is the predicted probability of heart disease (**target=1**) for a given **age**, **sex** if **fbs = a**.

2. G-computation (G-formula)

The G-formula version of the ATE is:

$$\sum_w \sum_y P(Y = y|A = 1, W = w)P(W = w) \; - \sum_w \sum_y P(Y = y|A = 0, W = w)P(W = w)$$

In our setting, conceptually we are going to:

- Fit an outcome model $Q(a, w) = E[Y \mid A = a, W = w]$ using **target** as outcome and predictors **fbs**, **age**, **sex**.
- Average predicted differences across the empirical distribution of **age**, **sex** in the sample.

Empirical (sample) G-computation estimator: $\widehat{\text{ATE}}_{\text{G-comp}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{Q}(1, W_i) - \hat{Q}(0, W_i)\}$.

Where:

- $\hat{Q}(1, W_i)$ is the predicted probability of **target=1** if individual i had **fbs=1** given their **age_i**, **sex_i**. and $\hat{Q}(0, W_i)$ is the analogous prediction for **fbs=0**.

3. AIPTW (Augmented Inverse Probability of Treatment Weighting)

Define: Outcome regression: $Q(a, w) = E[Y \mid A = a, W = w]$ and Propensity score: $g(a \mid w) = P(A = a \mid W = w)$ and in our dataset: $g(1 \mid W) = P(fbs = 1 \mid age, sex)$ and $g(0 \mid W) = 1 - g(1 \mid W)$. The AIPTW estimator of the ATE is:

$$\widehat{\text{ATE}}_{\text{AIPTW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(I\{A_i = 1\}/\hat{g}(1 \mid W_i)\left(Y_i - \hat{Q}(1, W_i)\right) + \hat{Q}(1, W_i)\right) - \left(I\{A_i = 0\}/\hat{g}(0 \mid W_i)\left(Y_i - \hat{Q}(0, W_i)\right) + \hat{Q}(0, W_i)\right)\right]$$

Intuition in our context:

- The terms $I\{A_i = a\}/\hat{g}(a \mid W_i)\left(Y_i - \hat{Q}(a, W_i)\right)$ are "residuals" reweighted by inverse probability of treatment.
- The $\hat{Q}(a, W_i)$ parts are the G-computation predictions.
- Combining them yields a doubly robust estimator: consistent if either the propensity model or the outcome model is correctly specified.

4. TMLE (Targeted Maximum Likelihood Estimation) for the ATE

TMLE uses the same equations:

- $Q(a, w)$: outcome regression and $g(a \mid w)$: propensity score

The targeted estimates are obtained as: $\widehat{\text{ATE}}_{\text{TMLE}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{Q}^*(1, W_i) - \hat{Q}^*(0, W_i)\}$

And: $\widehat{MOR}_{TMLE} = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(1, W_i)\right]\left[1 - \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(0, W_i)\right]}{\left[1 - \frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(1, W_i)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\hat{Q}^*(0, W_i)\right]}$

- Like in AIPTW, these estimators are also doubly robust.

2.7. Mediation Analysis

2.7.1. Motivations and Reasoning

Examining the DAG, it can be seen that there is one mediator variable, cholesterol, between the treatment(fbs) and outcome(target). Thus, mediation analysis must be conducted to determine the effect, if any, of the cholesterol level on the target variable. Three major confounders are considered in the analysis due to their interactions with fbs, cholesterol, and target. Age, a confounder on fbs and target, sex, a confounder on fbs and cholesterol, and ca, a confounder on cholesterol and the target variable. Since the study was not randomized, all three confounders must be considered.

The mediation analysis conducted is the natural linear parametric model. A natural parametric model was used as the mediator is not a variable that is manually set and there may be possible treatment-mediator interactions. This model assumes that all confounding between the three variables has been identified and adjusted.

2.7.2. Model Specifications

Model terms: $Y = Outcome(target), A = treatment(fbs), M = mediator(chol), H = confounders(age, sex, ca), M(a) =$ potential outcome of the mediator M to assigning treatment $A = a, Y(a, m) = potential\ outcome\ for\ the\ outcome\ Y\ to\ assigning\ treatment\ A = a\ and\ mediator\ M = m.$

Assumptions: (i) $Y(a, m) \perp\!\!\perp A|H$, (ii). $Y(a, m) \perp\!\!\perp M|H$, A, (iii) $M(a) \perp\!\!\perp A|H$, (iv) $Y(a, m) \perp\!\!\perp M(a)|H.$

Models

Natural direct effect(NDE): $NDE(a, a*;\ a*) = (\beta 1 + \beta 3(\alpha 0 + \alpha 1a + \alpha 2TE(H)))(a - a*)$

Natural indirect effect(NIE): $NIE(a, a*; a) = (\beta_2\alpha_1 + \beta_3\alpha_1)(a - a*)$

Total effect(TE): $TE = NDE + NIE$

Nonparametric bootstrapping of the results of NIE, NDE, and TE will be conducted to verify the validity of findings.

3. Results

3.1. Causal DAG and data exploration

The causal DAG was produced based on the review of literatures from previous studies using DAGitty, online version, and the result was given in figure 1

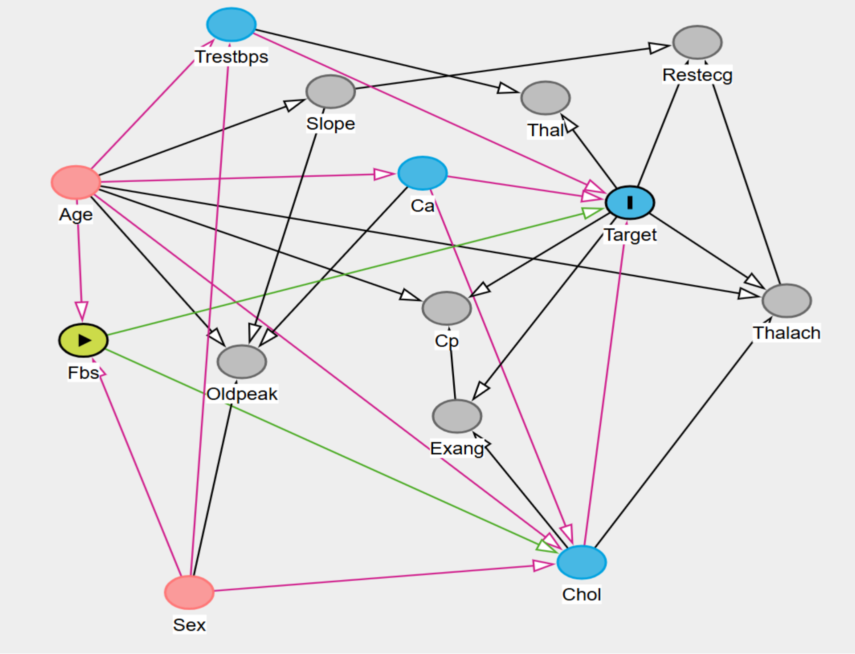
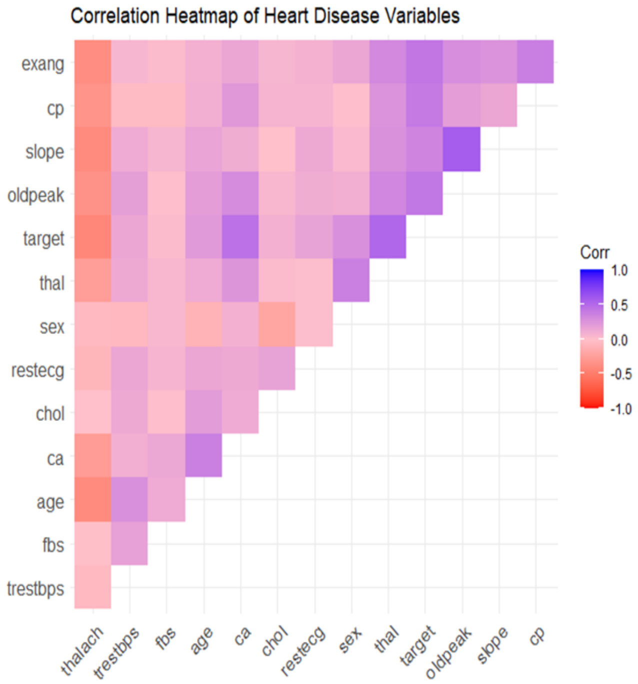


Figure 1: The DAG

Based on the DAG and review of the literature, we identified fbs as the exposure, age and sex to be confounders, cholesterol as a mediator, trestbps as an effect modifier, and target to be the outcome.

The correlation matrix and correlation matrix heatmap plot were used to understand the nature of relationship between variables, and check the validity of the initial DAGs, before proceeding to further analysis. From the result we can see that the outcome variable has non-negligible relationship with all variables included in the study, but the direction of relationship cannot be determined from the correlation analysis.



variables	***Corr***
Age	0.2231
Sex	0.2768
Cp	0.4144
Trestbps	0.1508
Chol	0.0852
Fbs	0.0253
Restecg	0.1692
Thalach	-0.4172
Exang	0.4319
Oldpeak	0.4245
Slope	0.3392
Ca	0.4600
Thal	0.5159
Target	1.0000

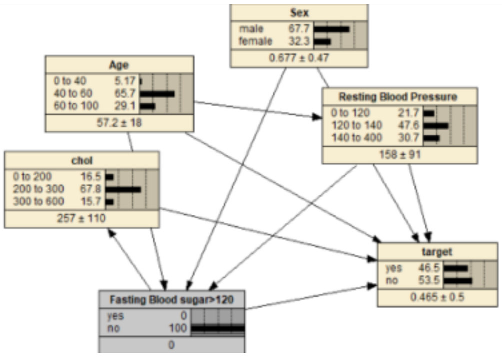
Figure 22: Exploratory Data analysis

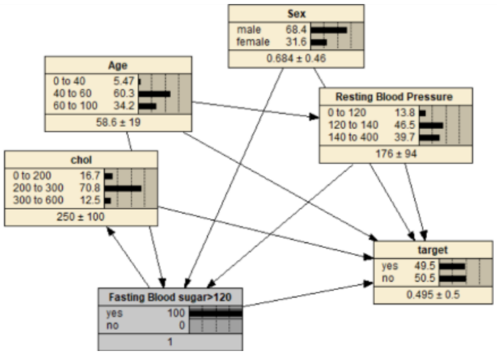
Numerical summary of target and exposure variable combination

Measure	Category	Count	Proportion
Exposure (fbs = a)	0	258	0.8515
Exposure (fbs = a)	1	45	0.1485
Outcome (y)	0	164	0.5413
Outcome (y)	1	139	0.4587
Crude Association	fbs = 0, y = 0	141	0.5465
Crude Association	fbs = 0, y = 1	117	0.4535
Crude Association	fbs = 1, y = 0	23	0.5111
Crude Association	fbs = 1, y = 1	22	0.4889

From the above results we see that the exposure-outcome variable combination is moderately balanced, so there is no extreme or unexpected outcome because of the imbalance exposure-outcome combination.

3.2. Results of Do-Calculus with Netica





$P(Y = 1|fbs = 1) = 0.495$ and

$P(Y = 1|fbs = 0) = 0.465 = \ ATE = 0.495 - 0.465 = 0.03$

So, from the results we see that the RD or ATE was 0.03 whereas the RR =1.0667 and OR=1.1277. Other details were omitted just to minimize the volume of our reports.

3.3. Results of Logistic regression

The results of fitting logistic regression to the data are given in table 2 below. The results indicate that the null deviance is very high as compared to the residual deviance, indicating that the treatment and confounders had done better in controlling the variability in Y, indicating the overall logistic regression model is better than the null model. The dispersion parameter was also found to be 1, indicating that there is no problem of overdispersion. The estimate of the coefficient of the treatment effect $\hat{\beta}_1 = 0.884$, with p-value of 0.723, suggesting that β_1 is nonsignificant, there is no significant effect of treatment on the outcome. Further, the analysis has revealed that the effect of both cofounders is significant. The result of model 2 indicates that the effects of mediator are nonsignificant at 5% level of significance but there was no significant change in the estimate of β_1 between the two models. We further checked the presence of multicollinearity using VIF and the result indicated that there was no problem of multicollinearity in the data set. The ROC procedure also revealed that the area under the curve is 0.7277 (the model is fair).

Table 2 Logistic regression output for model 1

Term	Estimate	std.error	Statistic	p.value	95% confidence Interval	
(Intercept)	0.008	0.900	-5.363	0.000	0.001	0.044

fbs>120	0.884	0.347	-0.355	0.723	0.445	1.747
Age	1.069	0.015	4.437	0.000	1.038	1.101
sexMale	4.506	0.290	5.187	0.000	2.588	8.100

```
Model diagnostics results
(Dispersion parameter for binomial family taken to be 1) Null deviance: 417.98 on 302
df: Residual deviance: 372.18  on 299  df:  AIC: 380.18
Goodness of fit test
Sum of squared errors     Expected value|H0        SD          Z           P
64.8367953             64.5614450            0.2570878  1.0710362  0.2841532
```

```
Pseudo.R.squared.for.model.vs.null
                              Pseudo.R.squared
 McFadden                             0.109580
 Cox and Snell (ML)                   0.140292
 Nagelkerke (Cragg and Uhler)         0.187484
Likelihood.ratio.test stat=45.802  pvalue=6.2477e-10
```

Table 3 Logistic Regression of outcome on exposure adjusted for confounders and mediator

Term	Estimate	Std.error	Statistic	P-value	95% CI	
(Intercept)	0.001	1.393	-5.377	0.000	0.000	0.008
Fbs>120	0.808	0.353	-0.605	0.545	0.401	1.613
Age	1.058	0.016	3.623	0.000	1.027	1.092
Sexmale	5.412	0.309	5.465	0.000	3.008	10.138
trestbps	1.015	0.008	1.926	0.054	1.000	1.031
Chol	1.005	0.003	1.787	0.074	1.000	1.010

	FBS	AGE	SEX	CHOL	TRESTBPS
VIF	1.04	1.14	1.18	1.11	1.11

```
Null deviance: 417.98  on 302  degrees of freedom:  Residual deviance: 364.70  on
297  degrees of freedom: AIC: 376.7
```

The logistic regression model at hand is then checked for its overall significance and then for its goodness of fit to the data and the results of all these procedures are given below table 2. From the result we can see that the logistic regression is significantly different from the null model, see also the Pseudo R square values which are fairly different from zero. We further did chi square goodness of fit test, with the result indicating that the test statistic is 1.0710362 with p-value of 0.2841532, suggesting that we do not have evidence against the null hypothesis. So, it is reasonable to fit logistic regression to model the data.

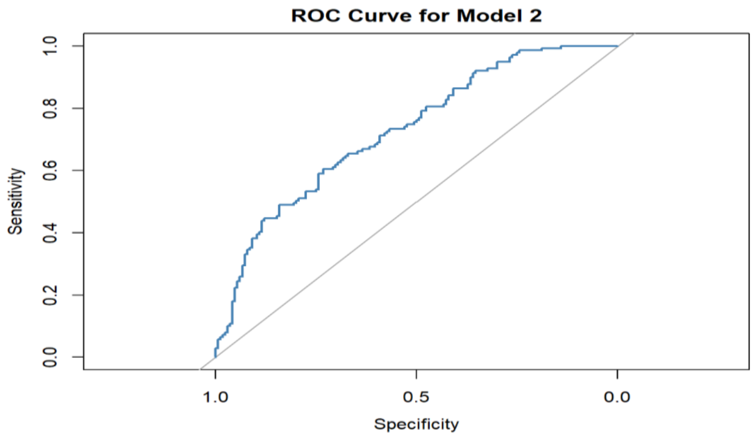


Figure 3:ROC of Logistic Regression Model

3.4. Bootstrap estimation procedure

The parameter estimate of the bootstrap procedure from simulation in stan was given in table 6. The result indicates that the posterior estimate of β_A was -0.19 with posterior high-density credible interval of (-0.91,0.51) with effective sample size of 3024 and $Rhat$ was 1.

Table 4 Posterior summary from Bayesian bootstrap of logistic regression parameters

Parameter	Mean	se_mean	Sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
Alpha	-7.01	0.03	1.34	-9.70	-7.93	-6.96	-6.08	-4.46	1792	**1**
beta_A	-0.19	0.01	0.36	-0.91	-0.44	-0.20	0.05	0.50	3024	**1**
beta_age	0.05	0.00	0.02	0.02	0.04	0.05	0.06	0.09	2096	**1**
beta_sex	1.64	0.01	0.30	1.08	1.44	1.64	1.85	2.22	2649	**1**
beta_chol	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	4364	**1**
beta_trestbps	0.01	0.00	0.01	0.00	0.01	0.01	0.02	0.03	2403	**1**

Samples were drawn using NUTS(diag_e) at Sun Nov 23 16:07:42 2025. For each parameter, n_eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat=1).

```
Bootstrap ATE:
the posterior estimate: Mean =-0.0406, SD = 0.0728,95%

CI = [-0.1657, 0.0786], total computation time 1.578047 hours
```

Posterior distribution of beta_1, details can be provided.

The sign of estimate of β_A negative indicates that treatment (increased fasting blood sugar level) has indirect effect on the risk of heart disease. The confidence interval includes 0, indicating that the effect is statistically nonsignificant. The $Rhat$ value of 1 indicates that simulation chain has converged. Finally, we observe from the result that the posterior estimate of the treatment effect ATE has mean of -0.0406, SD = 0.0728 with 95% credible interval of $[-0.1657, 0.0786]$,indicating the ATE under Bayesian modelling is also nonsignificant.

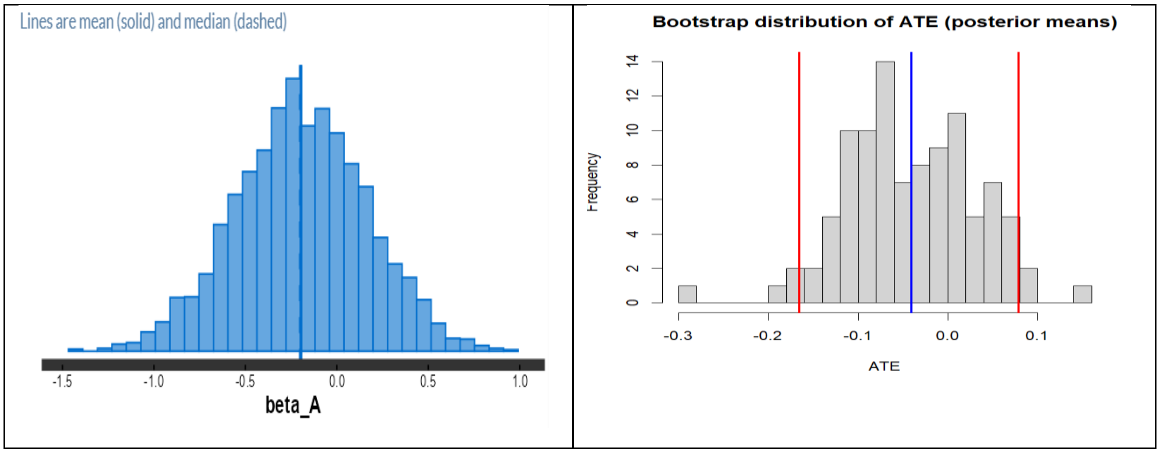


Figure 4: Posterior density of distribution of $\boldsymbol{\beta_A}$ and $\boldsymbol{ATE}$

3.5. Propensity Score Methods Results

The plots of propensity scores of treated and untreated groups were estimated using logistic regression, and given in figure 5. Because the density of the scores is not high near 0 or 1, the positivity assumption holds.

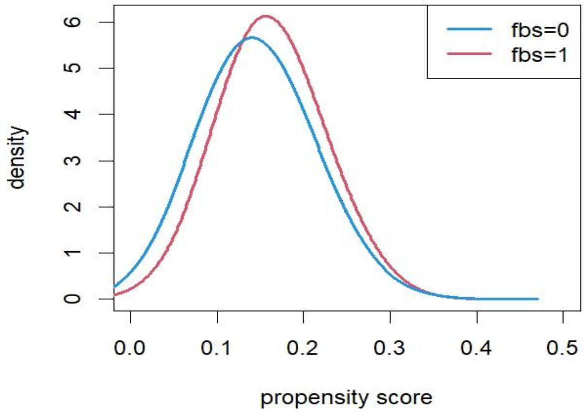


Figure 5: Density plot of propensity scores of treated and untreated groups

Logistic regression was used as described in the methodology, the exposure model is a weighted logistic regression with weights derived from propensity scores, and the outcome model is a logistic regression with fbs and the propensity scores. The table below shows the results for the weighted exposure model and outcome model. Both models show a causal OR near 1, and neither show a significant causal effect of high fasting blood sugar on heart disease.

Model	OR	Estimate	SE	Z Value	P Value	Significance
Exposure	1.07122	0.06880	0.1636	0.421	0.674	Not
Outcome	0.91960	-0.08382	0.3408	-0.246	0.806	Not

Using the "Matching" R package, matching was done on the basis of propensity scores, the estimated ATT is 0.040712, with p value 0.69537, again yielding a non-significant result for the ATT, meaning among the treated group, there was no significant causal effect of high fasting blood sugar on heart disease.

3.6. Targeted Maximum Likelihood Estimation, G-computation and AIPTW

This section summarizes the main estimated effects from the TMLE analysis of the treatment on the outcome, including marginal means, risk differences, and relative effect measures. The SuperLearner and tmle R packages were used to obtain these estimates.

Table 8: TMLE Estimate Effects

Quantity	*Estimate*	*95% CI (Lower, Upper)*	*p-value*
Marginal mean under treatment (EY1)	0.4732	0.3067, 0.6397	< 0.001
Marginal mean under control (EY0)	0.4626	0.4022, 0.5230	< 0.001
Additive effect (EY1 - EY0)	0.0106	-0.1663, 0.1875	0.906
Effect among the treated	-0.0281	-0.1925, 0.1363	0.738
Effect among the controls	0.0165	-0.1641, 0.1970	0.858
Relative risk (treatment vs. control)	1.0229	0.7031, 1.4881	0.906
Odds ratio (treatment vs. control)	1.0435	0.5131, 2.1223	0.906

The estimated average probability of the outcome if everyone received the treatment (EY1) is approximately 47%, whereas under control (EY0) it is about 46%. The additive treatment effect, defined as the difference in these marginal means, is very small (0.01) and not statistically significant (p = 0.906). This indicates no evidence that the treatment changes the overall probability of the outcome. Subgroup-specific effects among the treated and among the controls are likewise small and non-significant, with confidence intervals that also span zero. The relative risk (= 1.02) and odds ratio (= 1.04) for treated versus control individuals have wide confidence intervals including 1 and non-significant p-values, again suggesting no detectable treatment effect. Overall, across all effect measures (risk difference, relative risk, and odds ratio), the data do not support a statistically significant impact of the treatment on the outcome in this sample.

The table below summarizes the results of the G-computation, AIPTW, and TMLE analys**es.**

Table 9: G-computation, AIPTW and TMLE

Estimator	ATE Estimate	Std. Error	95% CI (Lower)	95% CI (Upper)
G-computation	-0.0262	0.000315	-0.0268	-0.0255
AIPTW	0.00439	0.088	-0.168	0.177
TMLE	-0.0213	0.09026	-0.161	0.118

We can see AIPTW and TMLE produce ATE estimates that are not statistically significant, as 0 is contained within both of their confidence intervals. G-computation shows a significant effect of the treatment on the outcome, although this effect is weak in magnitude. G-computation is not doubly robust and can be biased under model misspecification. Therefore, we will choose to believe that there is no ATE based on the results of this analysis.

3.7. Mediation Analysis Results

3.7.1. Parametric Linear Model

The results of the mediation analysis using the parametric linear modelling were given as NDE = -0.075 with confidence interval of (-0.212, 0.075), NIE = -0.003 with confidence interval of (-0.517, 0.031), and finally TE = -0.078 with confidence interval of (-0.220, 0.067). Confidence intervals are at 95% confidence and obtained through bootstrapping with 1,000 replicates. Within the context of the study, this would imply that a higher fasting blood sugar would reduce the likelihood of a heart disease diagnosis, which defies the initial assumptions obtained through research. These results would mean that the total effect of a higher fasting blood sugar lowers heart disease diagnosis by 7.76% altogether, and 7.46% without including the mediating effects of cholesterol. Since 0 is contained in all three confidence intervals, the results of the natural direct effect of fasting blood sugar on the target, the natural indirect effect of fasting blood sugar on the target (the effect obtained through cholesterol), and the total effect of fasting blood sugar on the target are inconclusive. It cannot be determined how much of an effect cholesterol has as a mediator on the effect of fasting blood sugar on heart disease. These results reflect the results obtained in the TMLE, G-computation, and AIPTW analyses in which no average treatment effect of fbs on target was found.

4. Conclusion

In this study, we have tried to establish causal relationship between heart disease and fasting blood sugar, adjusting for covariates and mediators in the Cleveland UCI heart disease data. Both graphical and numerical causal data analyses methods were used to establish the relationship. Dagitty and a correlation heatmap were used as graphical method of constructing causal relationships. Different causal data analysis methodologies were used, and the following results were obtained.

- The logistic regression of the outcome (heart disease) on the exposure (fasting blood sugar) adjusted for confounders and mediators were fitted, and the model was checked for its goodness and validities of the assumption for logistic regression and the result has revealed that the model fit to the data well but the effect the exposure on the outcome was found to be non-significant.
- The bootstrap estimate of the average treatment was done under the Bayesian set up, by imposing weak priors on the coefficient to obtain the prior estimate of the average treatment effect with its credible interval, and the posterior summary of regression coefficient. The results have indicated that the posterior estimate of ATE was not significantly different from zero and the posterior mean of the treatment effect adjusted for the covariate was also not significantly different from zero.
- Propensity score methods were used in the exposure model, outcome model and matching. Density of propensity score validate the positivity assumption. The weighted logistic regression exposure model estimates an OR greater than 1, but insignificant. The logistic regression outcome model estimates an OR less than 1, but insignificant. The matching method estimates a positive, but insignificant ATT. All three suggest the causal relationship of high fasting blood sugar on the presence of heart disease is not significant.
- TMLE, AIPTW, and G-Computation were run to estimate the average treatment effect of fbs on heart disease, adjusting for confounders of age and sex. For TMLE, it was found that both the additive effect and the odds ratio were non-significant. Also, the AIPTW estimate of the ATE was found to not be statistically significant. G-computation was significant, but this was disregarded as being a biased estimate due to model misspecification.
- Mediation analysis did not prove cholesterol to have any significant mediation effect between fasting blood sugar and the diagnosis of heart disease. Using the parametric linear model, insignificant results were found with the confidence interval for the normal indirect effect containing 0. It cannot be reasonably assumed that cholesterol has any mediation effect.

References

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Mensah, G. A., et al. (2023). Global Burden of Cardiovascular Diseases and Risks, 1990–2022. *Journal of the American College of Cardiology*.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41–55.

Seeman, T. E., et al. (1993). Health Behaviors, Stress, and Risk of Cardiovascular Disease. *American Journal of Epidemiology*.

Tervahauta, M., et al. (2000). Risk Factors of Coronary Heart Disease and Total Mortality among Elderly Men. *Age and Ageing*.

Yusuf, S., et al. (2004). Effect of Potentially Modifiable Risk Factors Associated with Myocardial Infarction in 52 Countries (INTERHEART Study). *The Lancet*.

Barone-Rochette, G., Piérard, S., De Meester, C., 2019. Diagnostic value of thallium myocardial perfusion imaging: A contemporary appraisal. *Journal of Nuclear Cardiology, 26*(1), 143–153.

Gulati, M., Levy, P. D., Mukherjee, D., et al. (2021). 2021 AHA/ACC chest pain guideline. *Circulation, 144*(22), e368–e454.

DeVon, H. A., Mirzaei, S., & Zegre-Hemsey, J. (2020). Typical and atypical symptoms of acute coronary syndrome: Time to retire the terms? *Journal of the American Heart Association, 9*(7).

Kim, K., Lee, H., & Park, J. (2023). Age-related risk stratification in cardiovascular disease. *Journal of Cardiology Research, 12*(2), 115–124.

Rodgers, J. L., Jones, J., Bolleddu, S. I., 2019. Cardiovascular risks associated with gender and aging. *Frontiers in Cardiovascular Medicine, 6*, 185.

Shahjehan, R. D., & El-Sherief, A. (2024). Coronary artery disease. In *StatPearls*. StatPearls Publishing.

Unger, T., Borghi, C., Charchar, F., et al. (2020). 2020 International Society of Hypertension Global Hypertension Practice Guidelines. *Hypertension, 75*(6), 1334–1357.

Appendix: R code used in the study

The following the main component of R code used in this project knitted from Rmakdown to word.

Causal Effect of Fasting Blood Sugar on Heart Attack

Group 3

Introduction

This report implements a logistic regression analysis to estimate the causal effect of fasting blood sugar (FBS) on heart attack (any heart disease) using the Cleveland UCI heart dataset. We fit two main models:

- Model 1: Adjusts for age and sex (total effect of FBS, conditional on these confounders).
- Model 2: Additionally adjusts for cholesterol and resting blood pressure (direct effect of FBS, controlling for potential mediators).

We present odds ratios and confidence intervals using `knitr::kable()` to create publication-ready tables.

Data Import

```
heart_raw <- readr::read_csv("C:/Users/16673/OneDrive - University of Missouri/
Fall 2025/Causal Inference/Project/Heart_disease_cleveland_new(in).csv", show_c
ol_types = FALSE)

heart_raw %>%

  head() %>%

  kable(caption = "Head of raw Cleveland dataset") %>%

  kable_styling(full_width = FALSE)
```

Head of raw Cleveland dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	0	145	233	1	2	150	0	2.3	2	0	2	0
67	1	3	160	286	0	2	108	1	1.5	1	3	1	1
67	1	3	120	229	0	2	129	1	2.6	1	2	3	1
37	1	2	130	250	0	0	187	0	3.5	2	0	1	0
41	0	1	130	204	0	2	172	0	1.4	0	0	1	0
56	1	1	120	236	0	0	178	0	0.8	0	0	1	0

Variable Definitions and Data Preparation

```
heart <- heart_raw %>%
  mutate(
    # Outcome: binary indicator of any heart disease
    target = ifelse(target > 0, 1, 0),
    target = factor(target, levels = c(0, 1), labels = c("No disease", "Disease")),

    # Exposure: FBS > 120 mg/dl (assumes original fbs is 0/1)
    fbs = factor(fbs, levels = c(0, 1), labels = c("<=120", ">120")),

    # Sex: 0 = female, 1 = male
    sex = factor(sex, levels = c(0, 1), labels = c("Female", "Male"))  )
# Keep complete cases for variables used in the main models
heart_cc <- heart %>%
  select(target, fbs, age, sex, chol, trestbps) %>%
  drop_na()
heart_cc %>%
  head() %>%
  kable(caption = "Head of analysis dataset (complete cases)") %>%
  kable_styling(full_width = FALSE)
```

Head of analysis dataset (complete cases)

target	fbs	age	sex	chol	trestbps
No disease	>120	63	Male	233	145
Disease	<=120	67	Male	286	160
Disease	<=120	67	Male	229	120
No disease	<=120	37	Male	250	130
No disease	<=120	41	Female	204	130
No disease	<=120	56	Male	236	120

Descriptive Statistics

```
# Summary of key variables
```

```
summary_table <- heart_cc %>%
  summarise(
    n = n(),
    mean_age = mean(age),
    sd_age = sd(age),
    mean_chol = mean(chol),
    sd_chol = sd(chol),
    mean_trestbps = mean(trestbps),
    sd_trestbps = sd(trestbps)
  ) %>%
  t() %>%
  as.data.frame() %>%
  rownames_to_column(var = "Statistic")
kable(summary_table, col.names = c("Statistic", "Value"),
      caption = "Basic descriptive statistics") %>%
  kable_styling(full_width = FALSE)
```

Basic descriptive statistics

Statistic	Value
n	303.000000
mean_age	54.438944
sd_age	9.038662
mean_chol	246.693069
sd_chol	51.776918
mean_trestbps	131.689769
sd_trestbps	17.599748

```
# Cross-tabulations
ftable_fbs_target <- table(heart_cc$fbs, heart_cc$target)
ftable_fbs_target %>%
  as.data.frame() %>%
  pivot_wider(names_from = Var2, values_from = Freq) %>%
  kable(caption = "FBS by heart disease status") %>%
  kable_styling(full_width = FALSE)
```

FBS by heart disease status

Var1	No disease	Disease
<=120	141	117
>120	23	22

Model 1: Total Effect of FBS (Adjusted for Age and Sex)

```
model1 <- glm(target ~ fbs + age + sex,
              data = heart_cc,family = binomial)

# Goodness of fit test
library(rms)

lrm_model <- lrm(target ~ fbs + age + sex, data = heart_cc,x = TRUE, y = TRUE)
residuals(lrm_model, "gof")
## Sum of squared errors     Expected value|H0                    SD
##            64.8367953            64.5614450             0.2570878
##                     Z                     P
##             1.0710362             0.2841532
library(rcompanion)

nagelkerke(model1)
## $Models
##
## Model: "glm, target ~ fbs + age + sex, binomial, heart_cc"
## Null:  "glm, target ~ 1, binomial, heart_cc"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                             0.109580
## Cox and Snell (ML)                   0.140292
## Nagelkerke (Cragg and Uhler)         0.187484
##
## $Likelihood.ratio.test
##   Df.diff LogLik.diff  Chisq    p.value
##        -3     -22.901 45.802 6.2477e-10
# $Number.of.observations
## Model: 303
## Null:  303
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting
 with ML"
##
## $Warnings
## [1] "None"
# Odds ratios and 95% CIs
model1_or <- tidy(model1, conf.int = TRUE, exponentiate = TRUE)


kable(model1_or,digits = 3,
      caption = "Model 1: Logistic regression of heart disease on FBS, age, and
 sex (odds ratios)") %>%
  kable_styling(full_width = FALSE)
```

Model 1: Logistic regression of heart disease on FBS, age, and sex (odds ratios)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.008	0.900	-5.363	0.000	0.001	0.044
fbs>120	0.884	0.347	-0.355	0.723	0.445	1.747
age	1.069	0.015	4.437	0.000	1.038	1.101
sexMale	4.506	0.290	5.187	0.000	2.588	8.100

Model 2: Direct Effect of FBS (Adjusted for Age, Sex, Cholesterol, and Resting BP)

```
model2 <- glm(target ~ fbs + age + sex + chol + trestbps,
              data = heart_cc,  family = binomial)


summary(model2)
##
## Call:
## glm(formula = target ~ fbs + age + sex + chol + trestbps, family = binomial,
##     data = heart_cc)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.489660   1.392869  -5.377 7.57e-08 ***
## fbs>120     -0.213751   0.353462  -0.605 0.545356
## age          0.056500   0.015595   3.623 0.000291 ***
## sexMale      1.688617   0.309006   5.465 4.64e-08 ***
## chol         0.004533   0.002536   1.787 0.073923 .
## trestbps     0.014849   0.007711   1.926 0.054131 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.98  on 302  degrees of freedom
## Residual deviance: 364.70  on 297  degrees of freedom
## AIC: 376.7
##
## Number of Fisher Scoring iterations: 3
library(rcompanion)
nagelkerke(model2)
## $Models
##
## Model: "glm, target ~ fbs + age + sex + chol + trestbps, binomial, heart_cc"
## Null:  "glm, target ~ 1, binomial, heart_cc"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                             0.127467
## Cox and Snell (ML)                   0.161247
## Nagelkerke (Cragg and Uhler)         0.215487
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq    p.value
##       -5      -26.64 53.279 2.9476e-10
##
```

```
## $Number.of.observations
##
## Model: 303
## Null:  303
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting
 with ML"
##
## $Warnings
## [1] "None"
model2_or <- tidy(model2, conf.int = TRUE, exponentiate = TRUE)


kable(model2_or,digits = 3,
      caption = "Model 2: Logistic regression adding cholesterol and resting BP
 (odds ratios)") %>%
  kable_styling(full_width = FALSE)
```

Model 2: Logistic regression adding cholesterol and resting BP (odds ratios)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.001	1.393	-5.377	0.000	0.000	0.008
fbs>120	0.808	0.353	-0.605	0.545	0.401	1.613
age	1.058	0.016	3.623	0.000	1.027	1.092
sexMale	5.412	0.309	5.465	0.000	3.008	10.138
chol	1.005	0.003	1.787	0.074	1.000	1.010
trestbps	1.015	0.008	1.926	0.054	1.000	1.031

Comparison of FBS Effect Across Models

```
comparison <- bind_rows(
  model1_or %>% mutate(model = "Model 1"),
  model2_or %>% mutate(model = "Model 2")
) %>%
  filter(term == "fbs>120") %>%
  select(model, estimate, conf.low, conf.high, p.value)

kable(comparison,
      digits = 3,
      caption = "Comparison of FBS effect between Model 1 and Model 2") %>%
  kable_styling(full_width = FALSE)
```

Comparison of FBS effect between Model 1 and Model 2

model	estimate	conf.low	conf.high	p.value
Model 1	0.884	0.445	1.747	0.723
Model 2	0.808	0.401	1.613	0.545

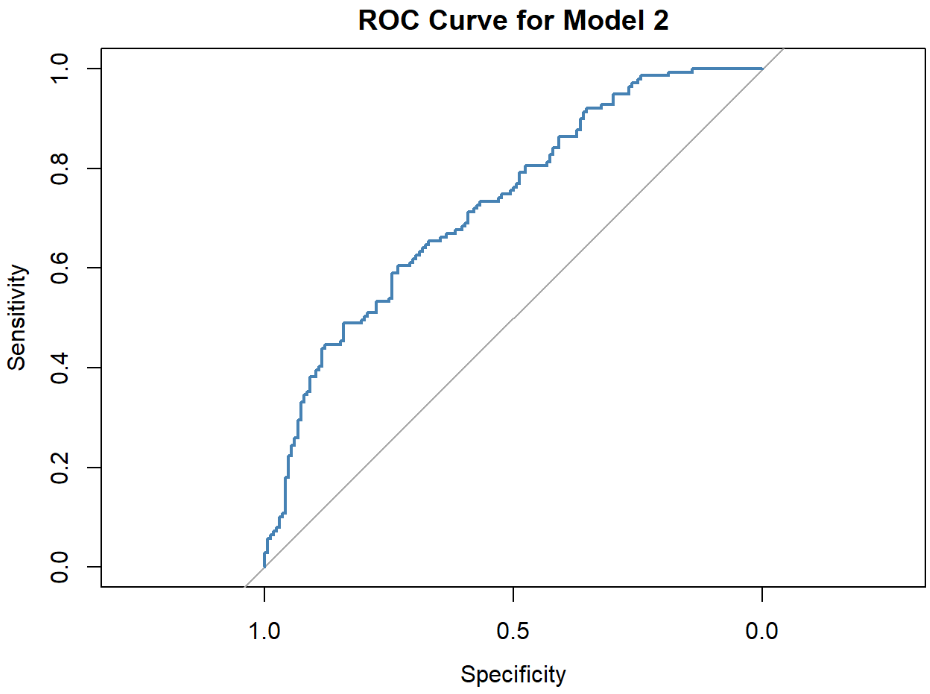
Model Diagnostics

```
# Hosmer-Lemeshow test for Model 2
hl_model2 <- ResourceSelection::hoslem.test(
  x = as.numeric(heart_cc$target) - 1,
  y = fitted(model2),  g = 10)

hl_model2
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  as.numeric(heart_cc$target) - 1, fitted(model2)
## X-squared = 9.1613, df = 8, p-value = 0.3289
# VIF for multicollinearity
vif_vals <- car::vif(model2)

vif_df <- tibble(
  term = names(vif_vals),
  VIF  = as.numeric(vif_vals))

kable(vif_df, digits = 2, caption = "Variance inflation factors (Model 2)") %>%
  kable_styling(full_width = FALSE)
```

Variance inflation factors (Model 2)

term	VIF
fbs	1.04
age	1.14
sex	1.18
chol	1.11
trestbps	1.11

```
# ROC curve and AUC for Model 2
roc_obj <- pROC::roc(heart_cc$target, fitted(model2))
auc_val <- pROC::auc(roc_obj)
auc_val
## Area under the curve: 0.7277
plot(roc_obj, col = "steelblue", main = "ROC Curve for Model 2")
```



Interpretation

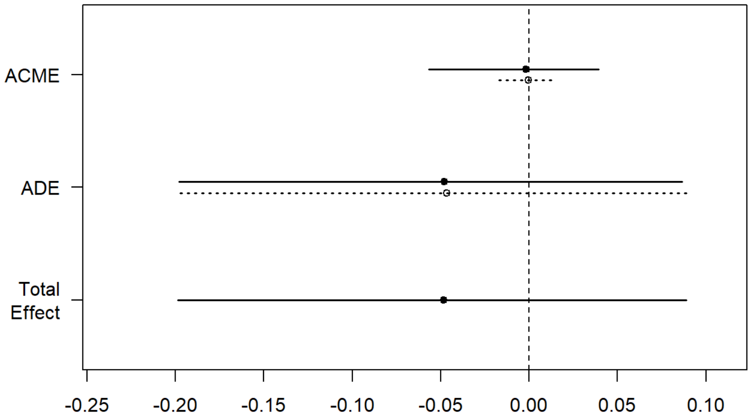
In Model 1, which adjusts only for age and sex, the odds ratio for FBS>120 represents the total effect of elevated fasting blood sugar on the odds of heart disease, conditional on these confounders.

Mediation using Package

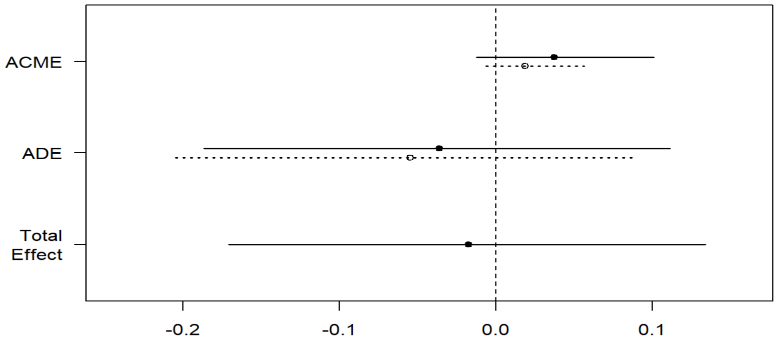
```
# R mediation script for Cleveland heart data
# Assumes: heart_cc data.frame with columns fbs, chol, trestbps, age, sex, targ
et


# --- Packages -----------------------------------------------------
if (!requireNamespace("mediation", quietly=TRUE)) install.packages("mediation")
if (!requireNamespace("boot", quietly=TRUE)) install.packages("boot")
if (!requireNamespace("dplyr", quietly=TRUE)) install.packages("dplyr")
library(mediation)
library(boot)
library(dplyr)


# --- Quick data checks ---------------------------------------------
# Drop obvious missing rows for this demonstration (or use multiple imputation
separately)
#heart_cc_cc <- heart_cc# %>% drop_na(fbs, chol, trestbps, age, sex, target)


# --- 1) Fit mediator models (two linear models, one per mediator) ------------
-


med_chol <- lm(chol ~ fbs + age + sex, data = heart_cc)
med_tbp  <- lm(trestbps ~ fbs + age + sex, data = heart_cc)


summary(med_chol)
##
## Call:
## lm(formula = chol ~ fbs + age + sex, data = heart_cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -129.62  -33.68   -4.29   28.74  289.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 200.6083    18.5762  10.799  < 2e-16 ***
## fbs>120      -0.6148     8.1518  -0.075 0.939934
## age           1.0987     0.3224   3.408 0.000745 ***
## sexMale     -20.0553     6.1994  -3.235 0.001353 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.01 on 299 degrees of freedom
## Multiple R-squared:  0.07622,    Adjusted R-squared:  0.06695
## F-statistic: 8.223 on 3 and 299 DF,  p-value: 2.825e-05
summary(med_tbp)
##
## Call:
## lm(formula = trestbps ~ fbs + age + sex, data = heart_cc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.505 -10.864  -1.376  10.265  62.161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 103.8960     6.2201  16.703  < 2e-16 ***
## fbs>120       7.2308     2.7296   2.649   0.0085 **
## age           0.5124     0.1080   4.746 3.22e-06 ***
## sexMale      -1.7247     2.0758  -0.831   0.4067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 299 degrees of freedom
## Multiple R-squared:  0.1036, Adjusted R-squared:  0.0946
## F-statistic: 11.52 on 3 and 299 DF,  p-value: 3.633e-07
# --- 2) Fit outcome model (binary outcome). Include interactions with each med
iator
# We use a logistic regression for target. Include product terms fbs*chol and f
bs*trestbps.


outcome_model <- glm(target ~ fbs * chol + fbs * trestbps + age + sex,
                     family = binomial, data = heart_cc)
summary(outcome_model)
```

```
##
## Call:
## glm(formula = target ~ fbs * chol + fbs * trestbps + age + sex,
##     family = binomial, data = heart_cc)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.002047   1.455579  -4.810 1.51e-06 ***
## fbs>120          -4.977487   3.275178  -1.520 0.128571
## chol              0.003206   0.002704   1.186 0.235772
## trestbps          0.012456   0.008617   1.445 0.148325
## age               0.058435   0.015667   3.730 0.000192 ***
## sexMale           1.755474   0.313735   5.595 2.20e-08 ***
## fbs>120:chol      0.011121   0.007504   1.482 0.138342
## fbs>120:trestbps  0.014677   0.019165   0.766 0.443785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.98  on 302  degrees of freedom
## Residual deviance: 361.69  on 295  degrees of freedom
## AIC: 377.69
##
## Number of Fisher Scoring iterations: 4
# ----------------------------------------------------------------------------
# Option A: Single-mediator mediation analysis using the 'mediation' package
# (run separately for chol and for trestbps). This estimates NIE and NDE on the
# outcome scale chosen by model (here logistic -> effects interpreted on probab
ility scale
# via averaging / quasi-Bayesian sims done by 'mediate').
#
# NOTE: mediation::mediate supports one mediator at a time.
# ------------------------------------------------------------------------------
--
# For mediator = chol
set.seed(2025)
```

```
med_out_chol <- mediate(model.m = med_chol,     # mediator model
                        model.y = outcome_model,# outcome model (must include m
ediator)
                        treat = "fbs",          # exposure
                        mediator = "chol",      # mediator name
                        boot = TRUE, sims = 1000) # sims for CIs
summary(med_out_chol)
##
## Causal Mediation Analysis
##
## Nonparametric Bootstrap Confidence Intervals with the Percentile Method
##
##                             Estimate 95% CI Lower 95% CI Upper p-value
## ACME (control)           -0.00041033  -0.01705061   0.01270288   0.878
## ACME (treated)           -0.00167265  -0.05677866   0.03934362   0.852
## ADE (control)            -0.04661810  -0.19727308   0.09104368   0.472
## ADE (treated)            -0.04788041  -0.19783817   0.08660507   0.432
## Total Effect             -0.04829074  -0.19870211   0.08905021   0.424
## Prop. Mediated (control)  0.00849711  -0.61115050   0.63561151   0.814
## Prop. Mediated (treated)  0.03463699  -2.38110791   3.30702741   0.804
## ACME (average)           -0.00104149  -0.03353458   0.02465882   0.864
## ADE (average)            -0.04724925  -0.19753319   0.08974822   0.462
## Prop. Mediated (average)  0.02156705  -1.51098649   1.84356713   0.804
## Sample Size Used: 303
## Simulations: 1000
plot(med_out_chol)
```



```
# For mediator = trestbps
set.seed(2025)
med_out_tbp <- mediate(model.m = med_tbp,
                       model.y = outcome_model,
                       treat = "fbs",
                       mediator = "trestbps",
                       boot = TRUE, sims = 1000)
summary(med_out_tbp)
##
## Causal Mediation Analysis
##
## Nonparametric Bootstrap Confidence Intervals with the Percentile Method
##
##                              Estimate 95% CI Lower 95% CI Upper p-value
## ACME (control)             0.0188068   -0.0063819    0.0563049   0.150
## ACME (treated)             0.0373535   -0.0122418    0.1010035   0.166
## ADE (control)             -0.0547907   -0.2047943    0.0905455   0.440
## ADE (treated)             -0.0362439   -0.1867432    0.1112281   0.612
## Total Effect              -0.0174372   -0.1708257    0.1339671   0.806
## Prop. Mediated (control) -1.0785458   -4.3125148    3.8268090   0.816
## Prop. Mediated (treated) -2.1421788   -8.3362744    8.3771470   0.912
## ACME (average)             0.0280801   -0.0017082    0.0702551   0.078 .
## ADE (average)             -0.0455173   -0.1941982    0.1017542   0.524
## Prop. Mediated (average) -1.6103623   -6.8294023    5.9285872    0.864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 303
## Simulations: 1000
plot(med_out_tbp)
```



```
dat <- read.csv("C:/Users/16673/OneDrive - University of Missouri/Fall 2025/Causal Inference/Project/Heart_disease_cleveland_new(in).csv")


m.mod <- glm(chol ~ fbs + age + sex + ca, data = dat)


y.mod <- glm(target ~ fbs * chol + age + sex + ca,
                data = dat, family = binomial)
summary(m.mod)
##
## Call:
## glm(formula = chol ~ fbs + age + sex + ca, data = dat)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 207.9085    19.2992  10.773  < 2e-16 ***
## fbs          -1.7194     8.1797  -0.210 0.833649
## age           0.9251     0.3460   2.674 0.007909 **
## sex         -21.1319     6.2401  -3.386 0.000803 ***
## ca            4.5884     3.3491   1.370 0.171713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2494.049)
##
##     Null deviance: 809616  on 302  degrees of freedom
## Residual deviance: 743227  on 298  degrees of freedom
## AIC: 3236.8
##
## Number of Fisher Scoring iterations: 2
summary(y.mod)
##
## Call:
## glm(formula = target ~ fbs * chol + age + sex + ca, family = binomial,
##     data = dat)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.369085   1.190757  -3.669 0.000243 ***
## fbs         -2.311391   1.956916  -1.181 0.237547
## chol         0.003274   0.002981   1.099 0.271956
## age          0.028222   0.017128   1.648 0.099411 .
## sex          1.682402   0.341785   4.922 8.55e-07 ***
## ca           1.188853   0.195514   6.081 1.20e-09 ***
## fbs:chol     0.007640   0.007764   0.984 0.325101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.98  on 302  degrees of freedom
## Residual deviance: 316.77  on 296  degrees of freedom
## AIC: 330.77
##
## Number of Fisher Scoring iterations: 4
natural_effects <- function(data, m.mod, y.mod){
  d <- data


  chol1 <- predict(m.mod, newdata = transform(d, fbs = 1), type = "response")
  chol0 <- predict(m.mod, newdata = transform(d, fbs = 0), type = "response")

  # Natural Direct Effect (NDE)
  # E[Y(fbs=1, M=M(0))] - E[Y(fbs=0, M=M(0))]

  y10 <- predict(y.mod,
                 newdata = transform(d, fbs = 1, chol = chol0),
                 type = "response")

  y00 <- predict(y.mod,
                 newdata = transform(d, fbs = 0, chol = chol0),
                type = "response")
    NDE <- mean(y10 - y00)
    #Natural Indirect Effect (NIE
  # E[Y(fbs=1, M=M(1))] - E[Y(fbs=1, M=M(0))]
    y11 <- predict(y.mod,
                 newdata = transform(d, fbs = 1, chol = chol1),
                 type = "response")
    y10b <- y10  # already computed
    NIE <- mean(y11 - y10b)
    #Total Effect
  TE <- mean(y11 - y00)
    list(NDE = NDE, NIE = NIE, TE = TE)
}
effects <- natural_effects(dat, m.mod, y.mod)
effects
## $NDE
## [1] -0.0745981
##
## $NIE
## [1] -0.003068275
##
## $TE
## [1] -0.07766638
library(boot)
boot_fun <- function(d, idx){
  dd <- d[idx, ]
  m.mod.b <- glm(chol ~ fbs + age + sex + ca, data = dd)
  y.mod.b <- glm(target ~ fbs * chol + age + sex + ca,
                 data = dd, family = binomial)
  ef <- natural_effects(dd, m.mod.b, y.mod.b)
  c(NDE = ef$NDE, NIE = ef$NIE, TE = ef$TE)
}
set.seed(123)
boot.out <- boot(data = dat, statistic = boot_fun, R = 1000)
boot.ci(boot.out, type = "perc", index = 1)  # NDE
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
## CALL :
## boot.ci(boot.out = boot.out, type = "perc", index = 1)
##
## Intervals :
## Level     Percentile
## 95%   (-0.2121,  0.0745 )
## Calculations and Intervals on Original Scale
boot.ci(boot.out, type = "perc", index = 2)  # NIE
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out, type = "perc", index = 2)
##
## Intervals :
## Level     Percentile
## 95%   (-0.0517,  0.0310 )
## Calculations and Intervals on Original Scale
boot.ci(boot.out, type = "perc", index = 3)  # TE
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
```

```
## CALL :
## boot.ci(boot.out = boot.out, type = "perc", index = 3)
##
## Intervals :
## Level     Percentile
## 95%   (-0.2198,  0.0666 )
## Calculations and Intervals on Original Scale
```