Estimating the Causal Effect of Sleep Apnea on Self-Reported General Health Using Targeted Maximum Likelihood Estimation in a Cross-Sectional Observational Study

Hamda Altaf[a], Rebekah Peter[a], Edward Zhao[a]
[a]Dalla Lana School of Public Health, University of Toronto

July 6, 2025

1 Introduction

Sleep apnea is a chronic sleep disorder, affecting 5% to 10% people worldwide, characterized by recurrent breathing interruptions during sleep[1]. It is increasingly recognized for its associations with a range of adverse health outcomes, including cardiovascular disease, daytime fatigue, and impaired cognitive functioning[2]. However, less is known about how a history of sleep apnea affects individuals' subjective assessments of their overall health. General health perception, often self-reported, is a robust predictor of healthcare utilization and future morbidity[3]. Investigating whether sleep apnea influences this perception can reveal important insights into the broader impact of sleep disorders on population health and inform strategies for screening and intervention[3,4].

While associations between sleep apnea and perceived health have been described, few studies have rigorously examined this relationship using formal causal inference approaches. In this study, we aim to estimate the causal effect of having ever been diagnosed with sleep apnea on the risk of reporting poor general health. Specifically, we ask: Does having sleep apnea increase the risk of poor general perception of one's health? To answer this question, we use observational data and apply causal inference methods to estimate the average treatment effect of a self-reported history of sleep apnea on self-rated general health status.

To support causal identification, we constructed a directed acyclic graph (DAG) that outlines the hypothesized relationships among the exposure (sleep apnea), the outcome (general health perception), and a set of observed and unobserved covariates. As shown in Figure 1, sleep apnea is assumed to have a causal effect on self-reported general health (GHP). However, both variables are also influenced by a set of measured pre-exposure covariates: age, sex, marital status, income, education, cigarette smoking, alcohol use, diabetes, and high blood pressure. These variables were selected based on prior literature and clinical knowledge as potential confounders, and are included in the adjustment set to block backdoor paths.

For example, individuals with chronic conditions like diabetes or hypertension may be more likely to develop sleep apnea and to report worse general health, independent of their apnea status. In addition to these observed covariates, the DAG includes unmeasured confounders such as access to healthcare, indicated by U, which may influence both the likelihood of receiving a sleep apnea diagnosis and self-perceived health.

Traditional regression methods estimate associations that may be confounded or biased in observational data. To obtain a causal estimate of the effect of sleep apnea on self-reported general health, we adopt targeted maximum likelihood estimation (TMLE), a doubly robust and semiparametrically efficient method for causal inference.

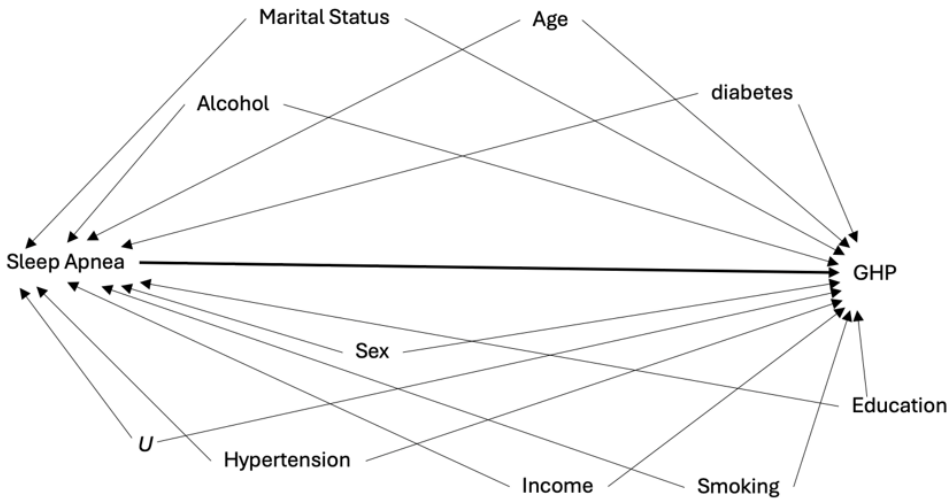


Figure 1: Causal diagram representing the effect of sleep apnea on self-reported general health (GHP). U represents unmeasured confounders, such as access to healthcare. To avoid clutter (and with no loss of validity of the backdoor criterion), the arrows between confounders have not been shown.

2 Methods

2.1 Data source

This study used a cross-sectional observational design, drawing on the CanPath Student Dataset, a synthetic, anonymized dataset developed by the Canadian Partnership for Tomorrow's Health (CanPath) to facilitate academic training and research[5]. This dataset was generated using the open-source R package synthpop, which produces synthetic versions of longitudinal health data by preserving the statistical relationships among variables while fully replacing and rearranging individual-level data. As such, the Student Dataset does not contain real participant information but reflects the structure and patterns of CanPath's nationally harmonized baseline and disease questionnaire data. The synthetic harmonized dataset includes: CanPath Baseline Questionnaire Data (2003-2017), CanPath Additional

Diseases Questionnaire Data (2003-2017), CanPath COVID-19 Questionnaire Data (2020), and CANOE Environmental Exposure Dataset (2006-2015).

The harmonized dataset includes over 40,000 observations and 403 categorical variables, encompassing a wide range of sociodemographic, economic, lifestyle, behavioral, and health-related information. Key domains include tobacco and alcohol use, nutrition, chronic disease diagnoses (e.g., diabetes, high blood pressure), and self-perceived health status. The dataset also incorporates environmental exposure variables from the Canadian Urban Environmental Health Research Consortium (CANUE), such as material deprivation indices and ambient air pollution measures. Access to the CanPath Student Dataset was granted for use in a university course, in accordance with CanPath's academic licensing agreement. The dataset is cross-sectional in structure and was designed to resemble CanPath's baseline and disease-focused questionnaires, while fully anonymizing individual-level records through synthetic data generation.

2.2 TMLE

Targeted maximum likelihood (TMLE) was used in this study to estimate the marginal causal effect of having sleep apnea on the risk of reporting poor self-rated general health. TMLE was selected because it produces a doubly robust estimator of the average treatment effect (ATE), remaining consistent as long as at least one of the models used to adjust for confounding or predict the outcome is correctly specified. In addition, TMLE allows for flexible machine-learning based model fitting, reducing the risk of model misspecification[6,7].

The analysis with TMLE was conducted using the `tmle` R package. The exposure was binary, defined as ever being diagnosed with sleep apnea before the age they were at the end of the questionnaire versus not being diagnosed. The outcome was also binary, defined as self-reporting fair or poor general health versus reporting good, very good, or excellent general health. Age, sex, cigarette smoking status, highest level of education completed, marital status, household income before taxes, ever consumed alcohol, ever diagnosed with diabetes, and ever diagnosed with high blood pressure were confounders that were adjusted for in the analysis.

To provide insight, the key steps executed internally by the `tmle` function are outlined here. The TMLE algorithm began by first fitting an outcome regression model (Q-model) with self-rated general health as the outcome and sleep apnea and confounders as predictors. Next, a separate model (g-model) was fit for the treatment assignment, where sleep apnea was the outcome and the confounders were the predictors. Both models were estimated using Super Learner libraries `SL.glm`, `SL.randomForest`, and `SL.gam` from the R package `SuperLearner`.

The estimated propensity scores (PS) from the g-model were then used to determine the clever covariate, H=A/PS-((1-A)/(1-PS)), where A represented the treatment assignment. Then, an intercept-free outcome regression model was fit to reduce residual bias, with self-reported general health as the outcome, the clever covariate as the predictor, and the predicted values from the Q-model as an offset. Lastly, the standardized risk of self-reported general health was estimated under both exposure scenarios: once by setting A equal to 1 for all individuals (diagnosed sleep apnea) and once by setting A equal to 0 (not diagnosed). The predicted risks across individuals were averaged to obtain the ATE. 95% confidence

intervals were determined using influence function based standard errors[7,8].

TMLE relies on the assumptions of consistency, positivity, and conditional exchangeability in causal inference to provide valid estimates[6]. This study assumes that consistency holds since exposure and outcome are well-defined binary indicators based on the data. The positivity assumption was validated in this study by visually examining the distribution of estimated propensity scores to ensure that no scores were close to 0 or 1 and that there was sufficient overlap between the exposed and unexposed groups. Individuals with propensity scores close to 0 or 1 as well as scores with insufficient overlap were excluded in order to satisfy the positivity assumption. Although the possibility of unmeasured confounding cannot be ruled out, the robustness of the findings to potential unmeasured confounding was assessed through a sensitivity analysis and by controlling for known confounders.

2.3 Sensitivity Analysis

The sensitivity analysis was conducted using the E-value, which is a measure of an association's robustness to potential uncontrolled confounders. The E-value minimum is 1 and so as the value gets farther from 1, the stronger the unmeasured confounding needs to be to explain away the observed association. If a confounder being associated by a risk ratio (RR) greater than the E-value is unlikely, then the association can be considered to be robust to unmeasured confounders. The E-value for the point estimate was calculated using the formula E-value $= RR^* + \sqrt{RR^* \times (RR^* - 1)}$,where $RR^* = max(RR, 1/RR)$[9]. The RR and odds ratio (OR) were estimated as secondary effect measures to support the sensitivity analysis using E-values and provide additional context.

2.4 Exploratory Subgroup Analysis

To explore potential heterogeneity in the effect of sleep apnea on general health perception, we conducted a subgroup analysis using causal forests. We fit a causal forest model with 2,000 trees, incorporating honesty and out-of-bag estimation to estimate conditional average treatment effects (CATEs) across individuals. As a preliminary step, we estimated the overall average treatment effect (ATE) to provide context. To improve the reliability of CATE estimates and address poor covariate overlap, we applied symmetric trimming by excluding individuals with estimated propensity scores below 0.1 or above 0.9[10].

To assess whether treatment effects varied systematically with baseline characteristics, we used best linear projection (BLP), regressing estimated CATEs on age, sex, education, income, smoking status, marital status, alcohol use, diabetes, and hypertension. This approach helped evaluate whether any covariates could be used to define subgroups with heterogeneous treatment effects.

2.5 Statistical Implementation and Missing Data

All analyses were conducted using R software (version 4.4.2, R Foundation for Statistical Computing)[11] . Analyses were performed using complete-case data, excluding individuals with missing values on the exposure, outcome, or any confounders. This study assumed that data are missing at random (MAR) conditional on the included covariates since a

comprehensive set of confounders were included in the analysis. This assumption supported the use of complete-case analysis.

3 Results

A comparison of baseline characteristics between individuals diagnosed with and without sleep apnea is presented in Table 1. We included 14730 patients in the analysis, 13904 (94.4%) with no sleep apnea diagnosis and 826 (5.6%) with sleep apnea diagnosis. After incorporating the variables from Table 1 into the treatment and outcome models, targeted maximum likelihood estimation (TMLE) was used to estimate the causal effect of sleep apnea on one's general perception of their health. Additionally, the covariate balance of the confounders was examined.

Table 1: Baseline characteristics stratified by sleep apnea status

Variable	**Category**	**Overall**	**Apnea = 0**	**Apnea = 1**	**p-value**
n		14730	13904	826	
Age (Mean (SD))		50.24 (11.27)	49.81 (11.23)	57.49 (9.17)	<0.001
Sex = 2 (%)		8586 (58.3)	8273 (59.5)	313 (37.9)	<0.001
Marital Status	Married	10712 (72.7)	10108 (72.7)	604 (73.1)	0.221
	Divorced	1326 (9.0)	1248 (9.0)	78 (9.4)	
	Widowed	416 (2.8)	385 (2.8)	31 (3.8)	
	Separated	681 (4.6)	642 (4.6)	39 (4.7)	
	Single	1595 (10.8)	1521 (10.9)	74 (9.0)	
Income	<$10,000	204 (1.4)	193 (1.4)	11 (1.3)	0.114
	<$24,999	934 (6.3)	885 (6.4)	49 (5.9)	
	<$49,999	2339 (15.9)	2195 (15.8)	144 (17.4)	
	<$74,999	2966 (20.1)	2781 (20.0)	185 (22.4)	
	<$99,999	2733 (18.6)	2581 (18.6)	152 (18.4)	
	<$149,999	3236 (22.0)	3056 (22.0)	180 (21.8)	
	<$199,999	1389 (9.4)	1335 (9.6)	54 (6.5)	
	>$200,999	929 (6.3)	878 (6.3)	51 (6.2)	
Education	None	22 (0.1)	20 (0.1)	2 (0.2)	<0.001
	Elementary	203 (1.4)	191 (1.4)	12 (1.5)	
	High School	2783 (18.9)	2606 (18.7)	177 (21.4)	
	Trade	1089 (7.4)	999 (7.2)	90 (10.9)	
	Diploma	3782 (25.7)	3610 (26.0)	172 (20.8)	
	Below Bachelor	687 (4.7)	638 (4.6)	49 (5.9)	
	Bachelor	3905 (26.5)	3715 (26.7)	190 (23.0)	
	Graduate	2259 (15.3)	2125 (15.3)	134 (16.2)	
Smoke Status	Never	7924 (53.8)	7519 (54.1)	405 (49.0)	0.037
	Past	4868 (33.0)	4567 (32.8)	301 (36.4)	
	Occasional	437 (3.0)	413 (3.0)	24 (2.9)	
	Daily	1501 (10.2)	1405 (10.1)	96 (11.6)	
Alcohol Use = 1 (%)		13746 (93.3)	12988 (93.4)	758 (91.8)	0.077
Diabetes = 1 (%)		1042 (7.1)	918 (6.6)	124 (15.0)	<0.001
Hypertension = 1 (%)		3580 (24.3)	3157 (22.7)	423 (51.2)	<0.001

Table 2: Estimated Causal Effect and Potential Impact of Sleep Apnea on Poor Self-rated General Health

Causal Effect/Potential Impact Measure	Estimate	95% CI	P value
ATE	0.0744	0.0422, 0.1066	<0.001
Risk Ratio	1.6347	1.3762, 1.9416	<0.001
Odds Ratio	1.7851	1.4446, 2.2059	<0.001

Table 2 shows the estimated marginal effect of sleep apnea on the risk of poor self-rated general health derived using TMLE. The estimated ATE of having sleep apnea was 0.0744. This indicates that having sleep apnea was associated on average with an 7.44% increase in the risk of reporting poor self-rated general health after adjusting for confounders. The 95% confidence interval for the ATE (0.0422, 0.1066) did not include 0, indicating that the effect was statistically significant.

Table 2 also presents the estimated risk ratio and odds ratio for additional context. The risk ratio was about 1.63, indicating that individuals with sleep apnea were about 1.63 times more likely to report poor general health than those without sleep apnea. The odds ratio was about 1.79, indicating that the odds of reporting poor general health was about 79% higher for individuals with sleep apnea compared to those without. Both the risk ratio and the odds ratio had 95% CIs that did not include 1, indicating that these effects were statistically significant.

The E-value for the estimate was calculated using $RR^* = RR$ since the RR was greater than 1. The E-value was equal to 2.653, indicating that an unmeasured confounder would need to be associated with both the treatment and outcome, conditional on other measured covariates, by a risk ratio greater than 2.653. This suggests that the observed association is moderately robust to unmeasured confounding.

After visually inspecting the distribution of the propensity scores, individuals with estimated propensity scores below 0.01 or above 0.25 were trimmed to satisfy the positivity assumption. This satisfied the positivity assumption by removing individuals with near-zero probability of receiving the exposure and helped ensure sufficient overlap between groups. 1227 individuals out of the 13904 from the unexposed group (no sleep apnea) were excluded and 14 individuals out of the 826 from the exposed group (sleep apnea) were excluded. The R code used to produce the distribution is provided in Appendix B.

To investigate heterogeneity in the effect of sleep apnea on general health, we used best linear projection (BLP) from the causal forest. None of the covariates showed statistically significant modification of the treatment effect (all $p > 0.4$), suggesting no strong evidence of heterogeneity across measured subgroups. However, small differences in treatment effects between subgroups may have gone undetected due to limited statistical power. For context, the overall average treatment effect (ATE) estimated from the causal forest was 0.08 (95% CI: 0.0445 to 0.1155), which decreased slightly to 0.0503 (95% CI: 0.0048 to 0.0958) after trimming individuals with extreme propensity scores to ensure overlap, suggesting a slightly attenuated but still positive association under conditions of good covariate overlap.Together, these findings support the robustness of the estimated association between sleep apnea and poor perceived health.

4 Discussion

In this study using TMLE, individuals who have ever received a sleep apnea diagnosis are, on average, 7.44% more likely to report poor general health than those without a sleep apnea diagnosis. Furthermore, the estimated risk ratio of 1.63 and odds ratio of 1.79 are evidence for sleep apnea as an independent risk factor of reporting poor general health. The marginal causal effect estimates of sleep apnea were adjusted for various sociodemographic

characteristics as well as smoking habits, lifetime alcohol use, diabetes, and hypertension as indicated by the causal diagram in Figure 1. The potential causal link between sleep apnea and poor general health perception may be explained by various physiological mechanisms including sleep fragmentation, intermittent hypoxia, and their effects on mood and energy levels[12,13,14]. To explore heterogeneity, we used causal forests to determine whether the effect of sleep apnea on self-rated general health varied across subgroups. We found no strong evidence of heterogeneity across measured subgroups, as none of the baseline covariates significantly modified the treatment effect. This supports the importance of addressing sleep apnea across diverse segments of the population, rather than targeting high-risk subgroups.

These results must be interpreted with limitations in mind. As with all causal analysis methods, we assume no unmeasured and unknown confounders. As previously mentioned, we adjusted for a wide range of known confounders, however the potential presence of unmeasured confounders may still lead to residual bias. Our calculated E-value is 2.653, which suggests that a relatively strong unmeasured confounder would be needed to fully explain away the observed effect, but it does not eliminate this possibility. Additionally, our analysis relied on the clinical diagnosis of sleep apnea, however, previous studies have shown that sleep apnea has been a largely under diagnosed condition[15]. This missclassification could introduce bias towards sleep apnea not being causative for reporting poor general health. Finally, there is temporal ambiguity in the analysis with some covariates. In our dataset, it is unclear whether marital status and alcohol use preceded or succeeded sleep apnea diagnosis. This complicates the directionality assumptions in the causal DAG and may introduce bias.

A strength of our analysis is TMLE uses an estimation procedure that remains valid as long as either the outcome model or the exposure model is correctly specified. However, if both models are misspecified, there is significant potential for bias. Another key advantage of TMLE is that it can incorporate machine learning methods while still preserving desirable statistical properties, such as consistency and asymptotic normality.

To summarize, those with a sleep apnea diagnosis may significantly increase the risk of reporting poor self-rated general health. This causal relationship suggests that intervening on sleep apnea could meaningfully improve people's subjective health. This is actionable from both a clinical and public health standpoint through proven therapies, such as continuous positive airway pressure (CPAP), that could improve how people perceive their health[16].

References

1. Hirani, R., & Smiley, A. (2023). A Scoping Review of Sleep Apnea: Where Do We Stand? Life, 13(2), Article 2. https://doi.org/10.3390/life13020387

2. Bjornsdottir, E., Keenan, B. T., Eysteinsdottir, B., Arnardottir, E. S., Janson, C., Gislason, T., Sigurdsson, J. F., Kuna, S. T., Pack, A. I., & Benediktsdottir, B. (2015). Quality of life among untreated sleep apnea patients compared to the general population and changes after treatment with positive airway pressure. Journal of Sleep Research, 24(3), 328–338. https://doi.org/10.1111/jsr.12262

3. Al-Windi, A., Dag, E., & Kurt, S. (2002). The influence of perceived well-being and reported symptoms on health care utilization: A population-based study. Journal of Clinical Epidemiology, 55(1), 60–66. https://doi.org/10.1016/S0895-4356(01)00423-1

4. Dalmases, M., Benítez, I., Sapiña-Beltran, E., Garcia-Codina, O., Medina-Bustos, A., Escarrabill, J., Saltó, E., Buysse, D. J., Plana, R. E., Sánchez-de-la-Torre, M., Barbé, F., & de Batlle, J. (2019). Impact of sleep health on self-perceived health status. Scientific Reports, 9(1), 7284. https://doi.org/10.1038/s41598-019-43873-5

5. CanPath—Canadian Partnership for Tomorrow's Health. (n.d.). CanPath - Canadian Partnership for Tomorrow's Health. Retrieved July 6, 2025, from https://canpath.ca/

6. Schuler, M. S., & Rose, S. (2017). Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American journal of epidemiology*, *185*(1), 65–73. https://doi.org/10.1093/aje/kww165

7. van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1), 1–38. https://doi.org/10.2202/1557-4679.1043

8. Abdollahpour, I., Nedjat, S., Almasi-Hashiani, A., Nazemipour, M., Mansournia, M. A., & Luque-Fernandez, M. A. (2021). Estimating the Marginal Causal Effect and Potential Impact of Waterpipe Smoking on Risk of Multiple Sclerosis Using the Targeted Maximum Likelihood Estimation Method: A Large, Population-Based Incident Case-Control Study. *American journal of epidemiology*, *190*(7), 1332–1340. https://doi.org/10.1093/aje/kwab036

9. VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of internal medicine*, *167*(4), 268–274.https://doi.org/10.7326/M16-2607

10. Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1), 187–199. https://doi.org/10.1093/biomet/asn055

11. R Core Team. (2024). R: A language and environment for statistical computing (Version 4.4.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

12. Puech, C., Badran, M., Runion, A. R., Barrow, M. B., Cataldo, K., & Gozal, D. (2023). Cognitive Impairments, Neuroinflammation and Blood–Brain Barrier Permeability in Mice Exposed to Chronic Sleep Fragmentation during the Daylight Period. International Journal of Molecular Sciences, 24(12), 9880. https://doi.org/10.3390/ijms24129880

13. Lavie L. (2015). Oxidative stress in obstructive sleep apnea and intermittent hypoxia–revisited–the bad ugly and good: implications to the heart and brain. Sleep medicine reviews, 20, 27–45. https://doi.org/10.1016/j.smrv.2014.07.003

14. Lavalle, S., Masiello, E., Iannella, G., Magliulo, G., Pace, A., Lechien, J. R., Calvo-Henriquez, C., Cocuzza, S., Parisi, F. M., Favier, V., Bahgat, A. Y., Cammaroto, G., La Via, L., Gagliano, C., Caranti, A., Vicini, C., & Maniaci, A. (2024). Unraveling the Complexities of Oxidative Stress and Inflammation Biomarkers in Obstructive Sleep Apnea Syndrome: A Comprehensive Review. Life (Basel, Switzerland), 14(4), 425. https://doi.org/10.3390/life14040425

15. Santilli, M., Manciocchi, E., D'Addazio, G., Di Maria, E., D'Attilio, M., Femminella, B., & Sinjari, B. (2021). Prevalence of Obstructive Sleep Apnea Syndrome: A Single-Center Retrospective Study. International journal of environmental research and public health, 18(19), 10277. https://doi.org/10.3390/ijerph181910277

16. Berg, L. M., Ankjell, T. K. S., Sun, Y. Q., Trovik, T. A., Rikardsen, O. G., Sjögren, A., Moen, K., Hellem, S., & Bugten, V. (2020). Health-Related Quality of Life and Sleep Quality after 12 Months of Treatment in Nonsevere Obstructive Sleep Apnea: A Randomized Clinical Trial with Continuous Positive Airway Pressure and Mandibular Advancement Splints. International journal of otolaryngology, 2020, 2856460. https://doi.org/10.1155/2020/2856460

A Appendix A: Summary of individual contributions

A.1 Edward Zhao

I contributed to the group report by first selecting covariates to be included in the analysis based on relevance to the causal question. I also wrote part of the results section, conducted the baseline descriptive analysis, and looked at baseline characteristics to compare potential confounders and assess the similarity of the sleep apnea and no sleep apnea groups. I also worked on the discussion section which included interpreting our main results and findings, acknowledging strengths and limitations of our study, such as the doubly robust property of TMLE and potential sources of bias, and presenting the public health relevance of our study.

A.2 Hamda Altaf

I contributed to the group report by conducting the background research and writing the introduction section. This included articulating the motivation for the study, clearly stating the study objective, and developing the conceptual framework through a Directed Acyclic Graph (DAG) to support causal identification. In the methods section, I described the study design and data source, including the structure and characteristics of the CanPath Student Dataset. I also conducted the subgroup analysis using causal forests, applying trimming procedures to address covariate overlap, and implemented best linear projection (BLP) to explore treatment effect heterogeneity. I wrote the corresponding section of the methods to describe this analytic approach in detail. In the results section, I interpreted and reported the findings from the subgroup analysis, including average and conditional treatment effects, and their implications for heterogeneity in the effect of sleep apnea on general health perception.

A.3 Rebekah Peter

I contributed to the group report by cleaning the source data and encoding the exposure and outcome variables in R. I conducted the TMLE analysis and sensitivity analysis in R, and wrote the corresponding sections of the methods (sections 2.2 and 2.3). I also wrote the methods section 2.5 to briefly explain the statistical implementation and missing data. I contributed to the results section by discussing the TMLE findings (ATE, odds ratio, risk ratio, and positivity assumption) and the sensitivity analysis results using the E-value.

Appendix B: R codes

```
library(causaldata)
library(tableone)
library(survey)
library(dplyr)

library(SuperLearner)
library(tmle)
library(gam)
library(randomForest)
library(ggplot2)

library(tableone)
library(xtable)

library(grf)
```

```
set.seed(123)

data_source<-read.csv("D:/Documents/Uoft_Biostat/Summer 2025/HAD7002H Causal Inference/final group project/CanPath_Student_Dataset_V2/CanPath_Student_Dataset_V2/student_dataset_canue_Version2_49900par_358var.csv")

#data cleaning
data_temp2<-data_source%>%
  filter((DIS_RESP_SLEEP_APNEA_EVER==0) | (DIS_RESP_SLEEP_APNEA_EVER==2) |
    (DIS_RESP_SLEEP_APNEA_EVER==1 & DIS_RESP_SLEEP_APNEA_AGE<SDC_AGE_CALC))


#View( data.frame(data_temp2$DIS_RESP_SLEEP_APNEA_EVER,data_temp2$SDC_AGE_CALC,data_temp2$DIS_RESP_SLEEP_APNEA_AGE))

data_temp2$apnea_ever<-NA
data_temp2$apnea_ever<-ifelse(data_temp2$DIS_RESP_SLEEP_APNEA_EVER==2,0,data_temp2$DIS_RESP_SLEEP_APNEA_EVER)

#View( data.frame(data_temp2$DIS_RESP_SLEEP_APNEA_EVER,data_temp2$apnea_ever,data_temp2$SDC_AGE_CALC,data_temp2$DIS_RESP_SLEEP_APNEA_AGE))

data_temp3<-data_temp2%>%
  filter(!is.na(HS_GEN_HEALTH))


data_temp3$health_stat<-NA
data_temp3$health_stat<-ifelse(data_temp3$HS_GEN_HEALTH==3|data_temp3$HS_GEN_HEALTH==4|data_temp3$HS_GEN_HEALTH==5,0,1)

#View(data.frame(data_temp3$HS_GEN_HEALTH,data_temp3$health_stat,data_temp3$DIS_RESP_SLEEP_APNEA_EVER,data_temp3$apnea_ever,data_temp3$SDC_AGE_CALC,data_temp3$DIS_RESP_SLEEP_APNEA_AGE))


data_final<-data.frame(SDC_AGE_CALC=data_temp3$SDC_AGE_CALC,
  SDC_SEX=data_temp3$SDC_SEX,
  SMK_CIG_STATUS=data_temp3$SMK_CIG_STATUS,
  SDC_EDU_LEVEL=data_temp3$SDC_EDU_LEVEL,
  SDC_MARITAL_STATUS=data_temp3$SDC_MARITAL_STATUS,
  SDC_INCOME=data_temp3$SDC_INCOME,
    ALC_EVER=data_temp3$ALC_EVER,
   DIS_DIAB_EVER=data_temp3$DIS_DIAB_EVER,
   DIS_HBP_EVER=data_temp3$DIS_HBP_EVER,
  health_stat=data_temp3$health_stat,
  apnea_ever=data_temp3$apnea_ever)

data_final$DIS_DIAB_EVER[data_final$DIS_DIAB_EVER == 2] <- 0
data_final$DIS_HBP_EVER[data_final$DIS_HBP_EVER == 2] <- 0


data_final<-na.omit(data_final)
```

```
set.seed(123)
#baseline table

covariates <- c(
  "SDC_AGE_CALC",
  "SDC_SEX",
  "SDC_MARITAL_STATUS",
  "SDC_INCOME",
  "SDC_EDU_LEVEL",
  "SMK_CIG_STATUS",
  "ALC_EVER",
  "DIS_DIAB_EVER",
  "DIS_HBP_EVER"
)

covFactors <- c(
  "SDC_SEX",
  "SDC_MARITAL_STATUS",
  "SDC_INCOME",
  "SDC_EDU_LEVEL",
  "SMK_CIG_STATUS",
  "ALC_EVER",
  "DIS_DIAB_EVER",
  "DIS_HBP_EVER")

table1 <- CreateTableOne(vars = covariates,
                         strata = "apnea_ever", data = data_final, factorVars = covFactors,
                         addOverall = TRUE)
table1
```

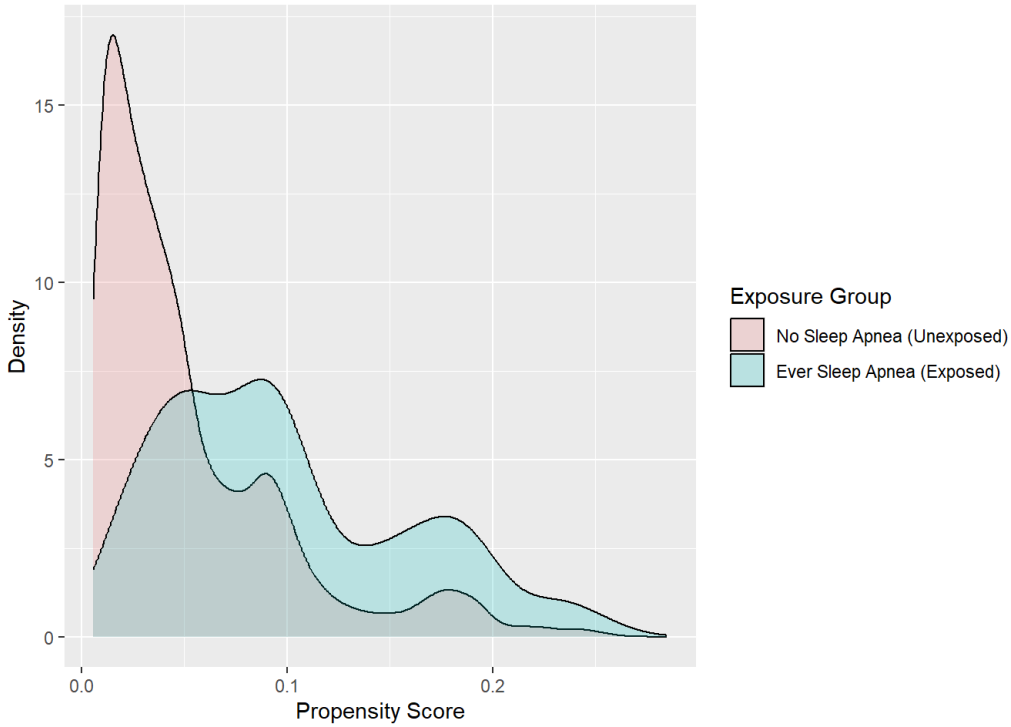
```
##                           Stratified by apnea_ever
##                            Overall       0             1             p
##   n                        14730         13904           826
##   SDC_AGE_CALC (mean (SD)) 50.24 (11.27) 49.81 (11.23) 57.49 (9.17)  <0.001
##   SDC_SEX = 2 (%)           8586 (58.3)   8273 (59.5)    313 (37.9)  <0.001
##   SDC_MARITAL_STATUS (%)                                              0.221
##      1                     10712 (72.7)  10108 (72.7)    604 (73.1)
##      2                      1326 ( 9.0)   1248 ( 9.0)     78 ( 9.4)
##      3                       416 ( 2.8)    385 ( 2.8)     31 ( 3.8)
##      4                       681 ( 4.6)    642 ( 4.6)     39 ( 4.7)
##      5                      1595 (10.8)   1521 (10.9)     74 ( 9.0)
##   SDC_INCOME (%)                                                     0.114
##      1                       204 ( 1.4)    193 ( 1.4)     11 ( 1.3)
##      2                       934 ( 6.3)    885 ( 6.4)     49 ( 5.9)
##      3                      2339 (15.9)   2195 (15.8)    144 (17.4)
##      4                      2966 (20.1)   2781 (20.0)    185 (22.4)
##      5                      2733 (18.6)   2581 (18.6)    152 (18.4)
##      6                      3236 (22.0)   3056 (22.0)    180 (21.8)
##      7                      1389 ( 9.4)   1335 ( 9.6)     54 ( 6.5)
##      8                       929 ( 6.3)    878 ( 6.3)     51 ( 6.2)
##   SDC_EDU_LEVEL (%)                                                  <0.001
##      0                        22 ( 0.1)     20 ( 0.1)      2 ( 0.2)
##      1                       203 ( 1.4)    191 ( 1.4)     12 ( 1.5)
##      2                      2783 (18.9)   2606 (18.7)    177 (21.4)
##      3                      1089 ( 7.4)    999 ( 7.2)     90 (10.9)
##      4                      3782 (25.7)   3610 (26.0)    172 (20.8)
##      5                       687 ( 4.7)    638 ( 4.6)     49 ( 5.9)
##      6                      3905 (26.5)   3715 (26.7)    190 (23.0)
##      7                      2259 (15.3)   2125 (15.3)    134 (16.2)
##   SMK_CIG_STATUS (%)                                                 0.037
##      0                      7924 (53.8)   7519 (54.1)    405 (49.0)
##      1                      4868 (33.0)   4567 (32.8)    301 (36.4)
##      2                       437 ( 3.0)    413 ( 3.0)     24 ( 2.9)
##      3                      1501 (10.2)   1405 (10.1)     96 (11.6)
##   ALC_EVER = 1 (%)         13746 (93.3)  12988 (93.4)    758 (91.8)   0.077
##   DIS_DIAB_EVER = 1 (%)     1042 ( 7.1)    918 ( 6.6)    124 (15.0)  <0.001
##   DIS_HBP_EVER = 1 (%)      3580 (24.3)   3157 (22.7)    423 (51.2)  <0.001
##                           Stratified by apnea_ever
##                            test
##   n
##   SDC_AGE_CALC (mean (SD))
##   SDC_SEX = 2 (%)
##   SDC_MARITAL_STATUS (%)
##      1
##      2
##      3
##      4
##      5
##   SDC_INCOME (%)
##      1
##      2
##      3
##      4
##      5
##      6
##      7
##      8
##   SDC_EDU_LEVEL (%)
##      0
##      1
##      2
##      3
##      4
##      5
##      6
##      7
##   SMK_CIG_STATUS (%)
##      0
```

```
##      1
##      2
##      3
##   ALC_EVER = 1 (%)
##   DIS_DIAB_EVER = 1 (%)
##   DIS_HBP_EVER = 1 (%)
```

```
set.seed(123)
#TMLE setting up outcome,exposure,covariates

#outcome
#1=poor, 0=good
data_Y<-data_final$health_stat

#exposure
#1=ever diagnosed with sleep apnea, 0=not
data_A<-data_final$apnea_ever

#covariates
data_W<-data_final[, c(
  "SDC_AGE_CALC",
  "SDC_SEX",
  "SMK_CIG_STATUS",
  "SDC_EDU_LEVEL",
  "SDC_MARITAL_STATUS",
  "SDC_INCOME",
    "ALC_EVER",
   "DIS_DIAB_EVER",
   "DIS_HBP_EVER"
)]

Q.SL.lib<-c("SL.glm","SL.randomForest","SL.gam")
g.SL.lib<-Q.SL.lib

#initial tmle to get PS
tmle.fit_initial <- tmle(Y=data_Y,
                 A=data_A,
                 W=data_W,
                 family="binomial",
                 V.Q=3, #outcome model;
                 V.g=3, #treatment model;
                 Q.SL.library=Q.SL.lib,
                 g.SL.library=g.SL.lib)
#check positivity and trim
prop_scores<-tmle.fit_initial$g$g1W
summary(prop_scores)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005571 0.019584 0.039281 0.056076 0.078386 0.284873
```

```
prop_score_df<-data.frame(propensity=prop_scores,
  exposure=factor(data_A,labels=c("No Sleep Apnea (Unexposed)","Ever Sleep Apnea (Exposed)")))

ggplot(prop_score_df,aes(x=propensity, fill=exposure))+
  geom_density(alpha=0.2)+ labs(main="Density of Estimated Propensity Scores by Exposure Group",x="Propensity Score",y="Dens
ity",fill="Exposure Group")
```




```
#trim based off of PS distribution
keep<-prop_scores>=0.01 & prop_scores <= 0.25

#trimmed dataset for analysis
data_final<-data_final[keep,]
data_Y<-data_Y[keep]
data_A<-data_A[keep]
data_W<-data_W[keep,]

#tmle for ate
tmle.fit <- tmle(Y=data_Y,
                 A=data_A,
                 W=data_W,
                 family="binomial",
                 V.Q=3, #outcome model;
                 V.g=3, #treatment model;
                 Q.SL.library=Q.SL.lib,
                 g.SL.library=g.SL.lib)
saveRDS(tmle.fit, file="SL_TMLE")
```

```
set.seed(123)
#tmle results
tmle.fit<-readRDS(file="SL_TMLE")
summary(tmle.fit)
```

```
##  Initial estimation of Q
##   Procedure: cv-SuperLearner, ensemble
##   Model:
##       Y ~  SL.glm_All + SL.randomForest_All + SL.gam_All
##
##   Coefficients:
##        SL.glm_All    0
##   SL.randomForest_All    0.2222188
##        SL.gam_All    0.7777812
##
##   Cross-validated pseudo R squared :  0.123
##
##  Estimation of g (treatment mechanism)
##   Procedure: SuperLearner, ensemble
##   Model:
##       A ~  SL.glm_All + SL.randomForest_All + SL.gam_All
##
##   Coefficients:
##        SL.glm_All    0
##   SL.randomForest_All    0
##        SL.gam_All    1
##
##  Estimation of g.Z (intermediate variable assignment mechanism)
##   Procedure: No intermediate variable
##
##  Estimation of g.Delta (missingness mechanism)
##   Procedure: No missingness, ensemble
##
##  Bounds on g: (0.0045, 1)
##
##  Bounds on g for ATT/ATC: (0.0045, 0.9955)
##
##  Marginal Mean under Treatment (EY1)
##    Parameter Estimate:  0.19163
##    Estimated Variance:  0.00026442
##               p-value:  <2e-16
##     95% Conf Interval:  (0.15975, 0.2235)
##
##  Marginal Mean under Comparator (EY0)
##    Parameter Estimate:  0.11723
##    Estimated Variance:  8.224e-06
##               p-value:  <2e-16
##     95% Conf Interval:  (0.11161, 0.12285)
##
##  Additive Effect
##    Parameter Estimate:  0.074399
##    Estimated Variance:  0.00027064
##               p-value:  6.1118e-06
##     95% Conf Interval:  (0.042156, 0.10664)
##
##  Additive Effect among the Treated
##    Parameter Estimate:  0.067285
##    Estimated Variance:  0.00019823
##               p-value:  1.7618e-06
##     95% Conf Interval:  (0.03969, 0.09488)
##
##  Additive Effect among the Controls
##    Parameter Estimate:  0.075791
##    Estimated Variance:  0.00028132
##               p-value:  6.2221e-06
##     95% Conf Interval:  (0.042917, 0.10866)
##
##  Relative Risk
##    Parameter  Estimate:  1.6347
##    Variance(log scale):  0.00771
##                p-value:  2.183e-08
##      95% Conf Interval:  (1.3762, 1.9416)
##
```

```
##  Odds Ratio
##     Parameter  Estimate:  1.7851
##     Variance(log scale):  0.011662
##                 p-value:  8.0507e-08
##       95% Conf Interval:  (1.4446, 2.2059)
```

```
cat("\n The ATE is",round(tmle.fit$estimates$ATE$psi,4)," and the 95% CI is (",round(tmle.fit$estimates$ATE$CI[1],4),",",rou
nd(tmle.fit$estimates$ATE$CI[2],4),")\n")
```

```
##
##  The ATE is 0.0744  and the 95% CI is ( 0.0422 , 0.1066 )
```

```
cat("\n The risk ratio is",round(tmle.fit$estimates$RR$psi,4)," and the 95% CI is (",round(tmle.fit$estimates$RR$CI[1],
4),",",round(tmle.fit$estimates$RR$CI[2],4),")\n")
```

```
##
##  The risk ratio is 1.6347  and the 95% CI is ( 1.3762 , 1.9416 )
```

```
cat("\n The odds ratio is",round(tmle.fit$estimates$OR$psi,4)," and the 95% CI is (",round(tmle.fit$estimates$OR$CI[1],
4),",",round(tmle.fit$estimates$OR$CI[2],4),")\n")
```

```
##
##  The odds ratio is 1.7851  and the 95% CI is ( 1.4446 , 2.2059 )
```

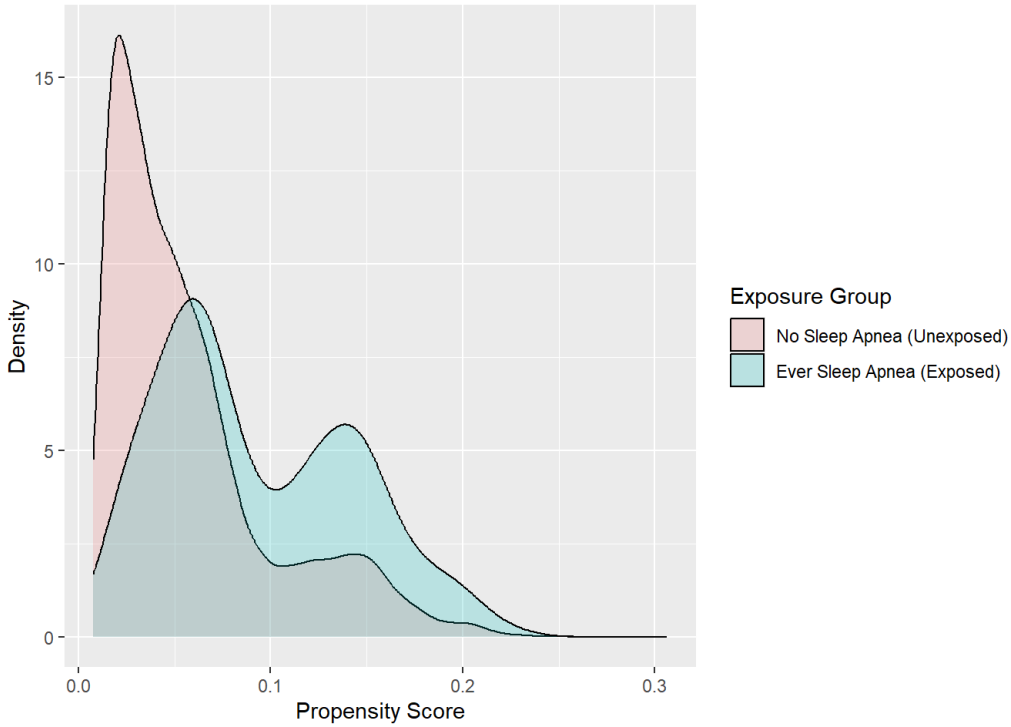
```
set.seed(123)
#e-value
rr_tmle<-tmle.fit$estimates$RR$psi
e_value<-rr_tmle+sqrt(rr_tmle*(rr_tmle-1))


cat("The E-value for the estimate is equal to",round(e_value,3),"\n.")
```

```
## The E-value for the estimate is equal to 2.653
## .
```

```
#ensuring positivity holds after final tmle
prop_scores<-tmle.fit$g$g1W

summary(prop_scores)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.007607 0.026374 0.046941 0.060197 0.076598 0.306414
```

```
prop_score_df<-data.frame(propensity=prop_scores,
  exposure=factor(data_A,labels=c("No Sleep Apnea (Unexposed)","Ever Sleep Apnea (Exposed)")))

ggplot(prop_score_df,aes(x=propensity, fill=exposure))+
  geom_density(alpha=0.2)+ labs(main="Density of Estimated Propensity Scores by Exposure Group",x="Propensity Score",y="Dens
ity",fill="Exposure Group")
```




```
#subgroup analysis
set.seed(123)
# Prepare data


X <- model.matrix(~ SDC_AGE_CALC + SDC_SEX + SDC_EDU_LEVEL + SDC_INCOME +
                    SMK_CIG_STATUS + SDC_MARITAL_STATUS + ALC_EVER +
                    DIS_DIAB_EVER+DIS_HBP_EVER- 1, data = data_final)

Y <- data_final$health_stat
W <- data_final$apnea_ever

# Fit causal forest
cf <- causal_forest(
  X,
  Y,
  W,
  num.trees = 2000,
  sample.fraction = 0.5,
  mtry = min(ceiling(sqrt(ncol(X)) + 20), ncol(X)),
  min.node.size = 10,
  honesty = TRUE,
  honesty.fraction = 0.5,
  honesty.prune.leaves = TRUE,
  alpha = 0.05,
  imbalance.penalty = 0,
  stabilize.splits = TRUE,
  ci.group.size = 2,
  tune.parameters = "none",
  compute.oob.predictions = TRUE
)
# Estimate average treatment effect
ate_cf <- average_treatment_effect(cf)
```

```
## Warning in average_treatment_effect(cf): Estimated treatment propensities go as
## low as 0.009 which means that treatment effects for some controls may not be
## well identified. In this case, using `target.sample=treated` may be helpful.
```

```
cat("ATE (Causal Forest):", round(ate_cf["estimate"], 4), "\n")
```

```
## ATE (Causal Forest): 0.08
```

```
cat("95% CI: (", round(ate_cf["estimate"]-1.96*ate_cf["std.err"], 4),",",round(ate_cf["estimate"]+1.96*ate_cf["std.err"],
4), ")\n")
```

```
## 95% CI: ( 0.0445 , 0.1155 )
```

```
set.seed(123)
# Step 1: Extract estimated propensity scores
prop_scores <- cf$W.hat

# Step 2: Filter to only units with good overlap
overlap_idx <- which(prop_scores > 0.1 & prop_scores < 0.9)

# Step 3: Subset the data
X_overlap <- X[overlap_idx, ]
Y_overlap <- Y[overlap_idx]
W_overlap <- W[overlap_idx]

data_overlap<-data_final[overlap_idx,]
# Step 4: Refit causal forest on trimmed sample
cf_overlap <- causal_forest(X_overlap, Y_overlap, W_overlap, num.trees = 2000)

# Step 5: Get trimmed ATE
ate_trimmed <- average_treatment_effect(cf_overlap)
cat("ATE (Trimmed sample):", round(ate_trimmed["estimate"], 4), "\n")
```

```
## ATE (Trimmed sample): 0.0503
```

```
cat("95% CI: (", round(ate_trimmed["estimate"]-1.96*ate_trimmed["std.err"], 4),",",round(ate_trimmed["estimate"]+1.96*ate_tr
immed["std.err"], 4), ")\n")
```

```
## 95% CI: ( 0.0048 , 0.0958 )
```

```
set.seed(123)
#best linear projection (blp)
#Subgroup effect summary
blp <- best_linear_projection(cf_overlap, X_overlap)
print(blp)
```

```
##
## Best linear projection of the conditional average treatment effect.
## Confidence intervals are cluster- and heteroskedasticity-robust (HC3):
##
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.0178864  0.3225352 -0.0555   0.9558
## SDC_AGE_CALC        0.0013180  0.0043501  0.3030   0.7619
## SDC_SEX             0.0049979  0.0586502  0.0852   0.9321
## SDC_EDU_LEVEL       0.0032981  0.0146598  0.2250   0.8220
## SDC_INCOME          0.0042665  0.0178142  0.2395   0.8107
## SMK_CIG_STATUS      0.0191721  0.0294493  0.6510   0.5151
## SDC_MARITAL_STATUS  0.0036006  0.0246264  0.1462   0.8838
## ALC_EVER           -0.0701327  0.0934472 -0.7505   0.4530
## DIS_DIAB_EVER       0.0329324  0.0717455  0.4590   0.6463
## DIS_HBP_EVER       -0.0204084  0.0582182 -0.3506   0.7260
```

```
# set.seed(123)
# #Explore CATEs by Subgroups if blp shows significant effect:
# cf_te <- predict(cf_overlap)$predictions
# data_overlap$cf_te <- cf_te
#
# ggplot(data_overlap, aes(x = factor(SDC_INCOME), y = cf_te)) +
#   geom_boxplot() +
#   labs(x = "Income Group", y = "Estimated Treatment Effect",
#        title = "Heterogeneous Treatment Effects by Income")
#
```