Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ 04/14/2025

Xiaxian Ou Date

Assessing Racial Disparities in Healthcare Expenditures Using Causal Path-Specific Effects

By

Xiaxian Ou
Master of Science in Public Health

Biostatistics

Razieh Nabi, Ph.D.
Committee Chair

David Benkeser, Ph.D.
Committee Member

Assessing Racial Disparities in Healthcare Expenditures Using Causal Path-Specific Effects

By

Xiaxian Ou
B.Med., Peking University, 2023

Committee Chair: Razieh Nabi, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2025

Abstract

Assessing Racial Disparities in Healthcare Expenditures Using Causal Path-Specific Effects
By Xiaxian Ou

Racial disparities in healthcare expenditures are well-documented, yet the underlying drivers remain complex and require further investigation. This study employs causal and counterfactual path-specific effects to quantify how various factors, including socioeconomic status, insurance access, health behaviors, and health status, mediate these disparities. Using data from the Medical Expenditures Panel Survey, we estimate how expenditures would differ under counterfactual scenarios in which the values of specific mediators were aligned across racial groups along selected causal pathways. A key challenge in this analysis is ensuring robustness against model misspecification while addressing the zero-inflation and right-skewness of healthcare expenditures. For reliable inference, we derive asymptotically linear estimators by integrating influence function-based techniques with flexible machine learning methods, including super learners and a two-part model tailored to the zero-inflated, right-skewed nature of healthcare expenditures.

Assessing Racial Disparities in Healthcare Expenditures Using Causal Path-Specific Effects

By

Xiaxian Ou
B.Med., Peking University, 2023

Thesis Committee Chair: Razieh Nabi, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2025

Acknowledgments

Looking back over the past two years, I am amazed at how much I have grown—both personally and academically. When I first left my family and home country in 2023 to begin my Master's studies in Biostatistics at Emory University, I was filled with uncertainty. Although I had discovered my interest in Biostatistics through previous projects, I initially lacked confidence due to my background in medicine, and I often questioned whether I could succeed in this new field. However, the exceptional support from the faculty, the thoughtfully designed curriculum, and the abundant resources offered by the program helped me build a strong foundation and gave me the courage to move forward. I am also deeply grateful to our program for providing me with the opportunity to continue my academic journey by pursuing a PhD.

I would like to express my deepest gratitude to my advisor, Dr. Razieh Nabi. I feel so fortunate to work with her. She is not only an outstanding researcher with strong academic expertise, but also a kind and patient mentor. Before working with Dr. Nabi, I was completely new to the field of causal inference. Her thoughtful and well-paced guidance helped me gradually understand and appreciate the depth and elegance of this area of research. Her strong theoretical background and her critical, rigorous thinking have deeply influenced me, encouraging me to approach research questions with depth and precision rather than relying on intuition alone. When I struggled with academic writing, Dr. Nabi provided detailed feedback on my manuscripts and shared exemplary work that helped me grow. I am also grateful for the many opportunities she provided to help me develop as a researcher, including participating in academic conferences and serving as a teaching assistant. Her generous support and strong recommendation played a crucial role in helping me secure a PhD offer at Emory. Her mentorship has been truly invaluable to me.

During my thesis project, I would also like to sincerely thank Anna Guo and Xinwei

He—two wonderful members of Dr. Nabi's lab. I still remember my first conversation with Anna about my research plan; her encouraging words reminded me not to set limits on myself, but to stay open-minded and just go for it. Her insightful blog posts and shared experiences were incredibly helpful in addressing some of the challenges in my research. I also truly enjoyed working with Xinwei on path-specific effect analysis, and I deeply appreciate her support, especially with the inference part of my thesis. I am grateful for their kindness, collaboration, and generosity throughout this journey.

During my two years at Emory, I would also like to express my sincere thanks to Dr. Xiangqin Cui, my APE advisor. Her extensive experience in collaborating with clinicians and her strong skills in interpreting results enhanced my ability to communicate with non-statistical audiences. I am also deeply grateful for her strong recommendation letter in my PhD application. Additionally, I truly appreciate the excellent coursework offered by the Biostatistics program at Emory, particularly the opportunity for master's students to take PhD-level classes. This academic flexibility provided me with a foundation to tackle complex problems in my research.

I would like to thank Lejun Wang and Zixuan Li—friends I made during my time at Emory. Together, we shared both the joys and the pressures of academic life, and their companionship has been a meaningful and comforting part of my journey. I also want to express my heartfelt thanks to my longtime friends Haonan Fu and Zihui Li in China. Despite the 12-hour time difference, we continued to share our stories, talk about our dreams, and support each other through life's challenges. Their friendship has been a constant source of strength and encouragement.

Then, I want to express my heartfelt thanks to my parents. Family has always been my safe harbor. In a traditional, ordinary Chinese family, giving up the opportunity to continue my master's studies at Peking University—which represented a stable and secure career path—and instead choosing to pursue a completely different direction

abroad, with high tuition and living expenses, is not an easy decision to accept. Yet, my parents gave me their full support. They respected my choice, gave me the freedom to follow my own path, and stood by me every step of the way. Their unwavering love and encouragement mean the world to me.

Finally, I want to thank myself for not giving up. I used to wonder: if there had been no accident in my college entrance examination major selection and spent those five years in a field I truly loved, would I become more competitive in an area? However, counterfactual outcomes can never be observed. The past cannot be changed—what matters is what we choose to do in the present moment. Over the past two years, I've also gradually come to realize that my medical training was not in vain. The intense workload helped me build endurance and cultivate resilience. I've also learned that it's never too late to start something new. I've come to accept that unexpected turns are a natural part of life. It's hard to always stay on the "right" track—especially in research—but we should never lose the courage to try. After all, the true value of climbing lies more in the journey itself than in the outcome.

Contents

List of Tables

1 Introduction

Racial disparities in health outcomes are well-documented public health challenges [32, 81]. Among these, disparities in healthcare expenditures are particularly consequential, reflecting inequities in access to and utilization of medical services [55]. Evidence from the Medical Expenditures Panel Survey (MEPS) consistently highlights these disparities in the United States [24, 52, 21, 73]. For instance, integrating MEPS data with the Medicare Current Beneficiary Survey, the National Health Interview Survey, and the Disease Expenditure project, Dieleman et al. [32] estimated that in 2016, White individuals—who comprised 61% of the U.S. population—accounted for 72% (95% uncertainty interval: 71%-73%) of total healthcare spending across all racial groups. This disparity underscores differential healthcare utilization between socially advantaged and marginalized populations, often reflecting avoidable and unjust inequities [15]. Understanding the mechanisms driving these disparities is essential for informing evidence-based strategies aimed at advancing equitable healthcare access. However, simple aggregate comparisons in healthcare spending overlook how disparities emerge through distinct pathways, necessitating a framework that isolates the contributions of different mediating factors.

Racial disparities in healthcare expenditures may arise from a complex interplay of socioeconomic, structural, and behavioral factors. *Socioeconomic status* (SES) is widely recognized as a primary driver, influencing access to resources, quality of care, and overall health outcomes [92]. Black and Hispanic populations, for instance, experience higher rates of poverty and lower levels of educational attainment compared to White populations [19], creating substantial barriers to affording and accessing healthcare [37]. *Access to insurance* further exacerbates these disparities, as uninsured or underinsured individuals are less likely to receive timely and adequate care [52, 46].

Zuvekas and Taliaferro [99] reported that health insurance accounted for 42% of the Black-White disparity and 24% of the Hispanic-White disparity in having a usual source of care. *Health behaviors*, which are shaped by cultural, social, and economic contexts, also play a role in shaping disparities [7]. Behavioral differences are evident across racial groups. For instance, smoking is associated with an annual increase of $1046 ($846-$1247) in per-capita healthcare expenses [1]. Studies also report that non-Hispanic Asian adults have the lowest prevalence of physical inactivity [20], while non-Hispanic White adults have the highest rate of cigarette smoking [18]. *Health status*, often shaped by cumulative disadvantages, further compounds disparities [53]. Minority populations report poorer self-rated health [9] and have higher rates of chronic conditions [47]. Despite higher expected medical spending burdens, these groups face greater barriers to healthcare and are often directed toward lower-quality, less comprehensive treatment [21]. Together, these factors create a complex web of influences driving racial disparities in healthcare expenditures. Understanding their mediating effects is crucial for identifying intervention points and informing policy solutions.

Empirical studies on racial disparities have traditionally relied on regression-based methods that compare outcomes or treatments across racial groups while adjusting for relevant covariates [89, 1, 91]. While informative, these approaches can be problematic when mediating factors are incorrectly treated as confounders, inadvertently blocking indirect pathways and failing to properly decompose racial disparities. To address these limitations, mediation analysis techniques, such as the Baron-Kenny approach, have been used to examine how disparities arise [8, 45, 12, 48, 31]. However, these methods often impose strong parametric assumptions, such as linearity, that may bias estimates when the underlying relationships involve non-linearities or interactions [70]. Moreover, traditional mediation analysis primarily decomposes effects into direct and indirect components, which can obscure the role of multiple mediators that operate

through distinct pathways.

To address limitations of traditional mediation approaches, we adopt a nonparametric counterfactual framework for path-specific effects (PSEs) [70, 85, 64]. PSEs allow us to quantify how race influences healthcare expenditures through distinct causal pathways, providing empirical insights for targeted policy interventions. For instance, if the PSE through insurance access is substantial, this suggests that differences in coverage contribute meaningfully to healthcare spending disparities. In this case, expanding Medicaid or implementing broader subsidies could help reduce inequities [61, 27]. Similarly, if the PSE through SES is large, policies focusing on education and income support may be more effective; if health behaviors play a key role, public health campaigns or tobacco taxation may help; if disparities are driven by health status, chronic disease management and preventive care should be prioritized. Using directed acyclic graphs [65], we formally map these pathways to quantify their contributions to racial disparities in healthcare expenditures. Our framework also avoids restrictive parametric assumptions by integrating flexible statistical and machine learning models into influence function-based estimators [84, 80, 83].

While causal mediation and path-specific effects provide powerful tools for analyzing disparities, the conceptualization of race as a causal variable remains contention in both causal inference and health disparities research [41, 88, 36, 86, 66, 39, 44]. As a socially constructed variable, race cannot be directly manipulated, like a traditional treatment variable in causal inference, challenging its causal interpretation under the principle of "no causation without manipulation" [41]. The total or mediated "effect of race" often reflects a composite of multiple dimensions—including physical phenotype, genetic background, and cultural context—inherently shaped by historical processes such as structural and institutional racism (e.g., Jim Crow laws and redlining). This complexity precludes the definition of plausible hypothetical interventions on race itself

[89, 43]. Researchers have proposed focusing on manipulable proxies, such as perceived race, to better understand the mechanisms driving racial disparities. This perspective aligns with the approach discussed in VanderWeele and Robinson [89], which advocates for estimating the extent to which racial inequality could be reduced through interventions on manipulable variables, such as insurance access. In line with these views, this study does not conceptualize race as a manipulable variable but rather as an analytical starting point to examine disparities in healthcare expenditures. Specifically, we assess how socioeconomic status, insurance access, health behaviors, and health status mediate racial disparities in healthcare expenditures, a structured framework for identifying interventions that could mitigate inequities. While race itself is not directly manipulable, targeting its key mediators—shaped by systemic racism and structural barriers–offers actionable pathways for reducing disparities. For example, policies that expand educational and economic opportunities, increase insurance coverage, promote healthier behaviors through public health initiatives, and improve chronic disease management can help mitigate inequities. Although such interventions have inherent limitations—since, for instance, manipulating SES may not fully capture the broader, nonmodifiable aspects of race—they still provide valuable insights into potential levers for change. By identifying mediated effects along specific pathways, our analysis highlights critical points for targeted interventions to address inequities in healthcare spending.

Beyond conceptual challenges, the estimation of path-specific effects presents several methodological challenges. Relationships between race, healthcare spending, and mediating factors are often complex and nonlinear, making model specification a key concern. Additionally, zero-inflation and right-skewness in expenditure data introduce further complications, requiring tailored statistical techniques. Existing methods—including plug-in G-computation [68, 95], inverse odds ratio-weighted estimators [78],

inverse treatment probability-weighted estimators [50], and regression-based imputation estimators [90, 98]—are widely used but prone to model misspecification bias. To mitigate these issues, we employ influence function-based estimators [84, 57, 30, 97], which provide some degree of robustness against model misspecification in parametric settings. A key advantage of these estimators, however, is their ability to accommodate data-adaptive statistical machine learning techniques, even when the underlying nuisance estimates converge at rates slower than parametric. Despite this flexibility, they still retain desirable frequentist properties, such as root-n consistency and asymptotic normality, which are crucial for constructing confidence intervals and quantifying uncertainty [22]. In our estimation pipeline, we employ super learners, which aggregate multiple predictive models to improve robustness and estimation accuracy while leveraging these statistical guarantees [67]. By integrating these techniques, our approach enhances the reliability of path-specific effect estimates, offering a more nuanced understanding of racial disparities in healthcare spending.

This study makes several key contributions to the literature on racial disparities in healthcare expenditures. First, we develop a path-specific effect framework to quantify the causal mechanisms driving racial differences in healthcare expenditures, offering a more granular and mechanistic perspective beyond traditional regression-based methods. Second, we advance estimation techniques by deriving asymptotically linear estimators based on influence function theory. We further integrate data-adaptive machine learning methods, such as super learners, to enhance estimation precision, improve robustness against model misspecification, and effectively handle the complex data-generating mechanisms underlying healthcare expenditures. Third, we apply this framework to analyze key mediators—including socioeconomic status, insurance access, health behaviors, and health status—using the 2009 and 2016 MEPS data, providing empirical insights into the pathways through which disparities arise. Finally, we

contribute the flexPaths R package, expanding the methodological toolkit for causal path-specific analysis in the study of racial disparities and beyond.

2 MEPS data and sample description

2.1 Data Source

The Medical Expenditures Panel Survey (MEPS), co-sponsored by the Agency for Healthcare Research and Quality and the National Center for Health Statistics, is a large-scale survey that collects detailed data on healthcare costs, use, and insurance coverage from families, individuals, medical providers, and employers across the United States. MEPS is a crucial resource for health services research and policy analysis due to its comprehensive individual-level data. For our analysis, we used the MEPS household components of the 2009 and 2016. The sample size for 2009 MEPS data was 20,816 after focusing on self-reported non-Hispanic Whites (9,963), non-Hispanic Blacks (3,971), Asians (1,469), and Hispanics (5,413). The 2016 MEPS data included 19,529 participants, consisting of self-reported non-Hispanic Whites (8,772), non-Hispanic Blacks (3,584), Asians (1,537), and Hispanics (5,636).

2.2 Variables

The MEPS samples collected information on individuals' baseline characteristics, SES, health insurance access, health behaviors, health status, and healthcare expenditures across different racial groups. A detailed breakdown of these variables is provided below.

Baseline characteristics include demographic information such as age and sex, as well as geographic region. Age is recorded as the exact age of each individual as of December 31 of the survey year, with the sample ranging from 18 to 85 years old. Sex, which includes male and female, was verified and corrected during each MEPS interview. Geographic region is categorized according to U.S. Census regions:

Northeast, Midwest, South, and West.

SES was measured by income and education. Income level was computed by dividing family income by the applicable poverty line (based on family size and composition) and classified into one of five categories: negative or poor (less than 100%), near poor (100% to less than 125%), low income (125% to less than 200%), middle income (200% to less than 400%), and high income (greater than or equal to 400% of the poverty line). Education was categorized into four levels: less than high school, high school, college, and graduate education.

For *insurance access*, individuals were considered uninsured if they were not covered by one of the following sources in the survey year: TRICARE, Medicare, Medicaid, SCHIP, or other public hospital/physician insurance, or private hospital/physician insurance.

Health behaviors were assessed using two variables: smoking and exercise. Smoking status indicated whether an individual was a current smoker, while exercise indicated whether a person had currently spent half hour or more in moderate to vigorous physical activity at least five times a week.

Health status was measured across several dimensions: (1) anthropometric measures, such as BMI (kg/m^2); (2) health perception, including perceived health status and perceived mental health status (both measured on a 5-point scale: excellent, very good, good, fair, and poor), as well as Physical Component Summary (PCS) and Mental Component Summary (MCS) scores; (3) functional status, assessed by cognition limitations, social limitations (such as the use of assistive technology and recreation), and any limitations in daily living activities, functional, or sensory abilities; and (4) chronic conditions, including diabetes, asthma, high blood pressure, coronary heart disease, angina, myocardial infarction, stroke, emphysema, cholesterol, arthritis, and cancer.

The *outcome* of interest is annual total healthcare expenditures, defined as the sum of direct payments for care provided during the year, including out-of-pocket payments and payments by private insurance, Medicaid, Medicare, and other sources. Payments for over-the-counter drugs are not included in MEPS total expenditures.

2.3 Sample description

Table 2.1 presents descriptive statistics on baseline characteristics, SES, insurance access, health behaviors, health status, and healthcare expenditures across the four racial groups in both 2009 and 2016. The racial composition was similar between 2009 and 2016, with non-Hispanic Whites comprising approximately half of the overall sample, while Asians accounted for the smallest proportion, around 7%. Whites had the highest median healthcare expenditures at 1,675 $ in 2009 and at 2,093 $ in 2016 repsectively, whereas Hispanics reported the lowest median expenditures during the same periods. The medians of healthcare expenditures increased across all racial groups from 2009 to 2016. To assess whether various factors differed significantly across the racial groups, categorical variables were compared across racial groups using the Chi-square test, while continuous variables were compared using Kruskal-Wallis rank sum test. Significant differences in SES, insurance access, health behaviors, and health status were observed across all racial groups within 2009 and 2016.

Table 2.2 shows the median healthcare expenditures in both 2009 and 2016 stratified by race and other characteristic levels. Overall, older adults and those living in northern and midwest regions tended to have higher median expenditures. Females spent more in healthcare compared with males. Additionally, individuals with higher educational attainment and income levels, as well as those enrolled in insurance programs, had significantly higher healthcare expenditures — nearly 1,400 $ difference of median for the insured compared to the uninsured. Conversely, participants who engaged in

regular exercise and reported better health status had lower healthcare expenditures. These expenditure trends were consistent across the four racial groups.

Table 2.1: Characteristics across different racial groups

	MEPS data in year 2009					MEPS data in year 2016				
Characteristic	Overall	Asians	Blacks	Hispanics	Whites	Overall	Asians	Blacks	Hispanics	Whites
N	20,816	1,469	3,971	5,413	9,963	19,529	1,537	3,584	5,636	8,772
Expenditure	920.0	540.0	758.0	283.0	1,675.0	1,118.0	777.0	888.5	396.0	2,093.0
Expenditure > 0 (%)	81.0%	80.4%	79.8%	67.2%	89.0%	81.9%	82.3%	78.4%	70.6%	90.6%
baseline characteristics										
Age	44.0	43.0	44.0	39.0	48.0	46.0	44.0	46.0	41.0	52.0
Male	45.6%	46.8%	40.2%	46.8%	46.9%	45.9%	47.4%	41.5%	45.9%	47.3%
Region										
North	15.0%	14.8%	17.1%	13.5%	15.1%	16.1%	15.7%	16.7%	14.9%	16.7%
Midwest	20.0%	10.8%	16.1%	10.1%	28.3%	19.4%	12.0%	16.4%	8.7%	28.7%
South	38.3%	17.2%	58.5%	34.3%	35.6%	38.4%	20.4%	57.7%	38.4%	33.7%
West	26.6%	57.2%	8.3%	42.0%	21.0%	26.1%	51.9%	9.1%	37.9%	20.9%
SES										
Income										
Below poverty	17.2%	9.9%	25.4%	24.0%	11.3%	17.3%	9.6%	26.0%	23.8%	11.0%
Near poverty	5.5%	2.9%	6.6%	7.5%	4.5%	5.4%	4.8%	6.6%	7.6%	3.6%
Low	16.3%	13.3%	18.4%	22.0%	12.7%	15.6%	11.8%	17.3%	20.9%	12.1%
Middle	31.1%	29.1%	30.2%	31.9%	31.4%	29.2%	23.5%	29.3%	31.2%	29.0%
High	29.9%	44.8%	19.4%	14.7%	40.1%	32.4%	50.4%	20.8%	16.6%	44.3%
Education										
< High school	26.5%	14.4%	26.4%	49.3%	15.8%	23.6%	13.1%	22.0%	42.8%	13.9%
High school	44.4%	30.9%	51.9%	37.3%	47.3%	42.8%	29.1%	53.9%	39.2%	43.0%
College	14.7%	31.0%	10.1%	6.9%	18.3%	16.4%	30.3%	10.5%	9.3%	21.1%
Graduate	14.5%	23.8%	11.6%	6.5%	18.5%	17.1%	27.5%	13.6%	8.7%	22.0%
Insurance access										
Uninsured	20.2%	14.0%	18.5%	38.5%	11.8%	12.0%	5.5%	10.0%	25.4%	5.3%
Health behaviors										
Smoke	18.1%	8.8%	21.5%	12.1%	21.4%	14.1%	7.3%	19.5%	8.9%	16.4%
Exercise	56.6%	58.9%	53.1%	52.5%	59.9%	49.6%	45.2%	50.9%	46.4%	51.8%
Health status										
BMI	27.1	23.7	28.3	27.5	26.6	27.4	24.1	29.0	28.2	27.1
Mental health										
Excellent	36.3%	42.3%	37.4%	34.8%	35.7%	35.2%	40.1%	38.2%	36.1%	32.5%
Very good	29.4%	29.6%	25.5%	28.1%	31.6%	28.7%	30.3%	25.1%	24.4%	32.8%
Good	26.5%	23.4%	27.5%	29.5%	24.9%	26.9%	23.0%	27.1%	30.2%	25.4%
Fair	6.3%	3.3%	7.8%	6.6%	6.1%	7.4%	5.3%	7.8%	8.1%	7.1%
Poor	1.5%	1.4%	1.8%	0.9%	1.7%	1.8%	1.2%	1.8%	1.2%	2.2%
Health										
Excellent	23.4%	26.8%	21.5%	21.3%	24.7%	23.1%	27.9%	22.3%	23.9%	22.1%
Very good	31.5%	34.4%	28.4%	28.0%	34.2%	31.9%	36.2%	28.3%	26.0%	36.3%
Good	30.1%	28.9%	31.8%	33.7%	27.7%	29.5%	27.3%	31.3%	32.3%	27.4%
Fair	11.6%	7.7%	14.5%	14.0%	9.6%	12.3%	6.5%	14.6%	15.1%	10.5%
Poor	3.5%	2.2%	3.8%	2.9%	3.8%	3.3%	2.1%	3.5%	2.7%	3.7%
PCS	53.2	54.2	52.1	53.7	52.9	53.5	54.8	52.6	53.8	53.2
MCS	53.0	54.0	53.3	51.7	53.7	54.4	54.9	54.8	54.4	54.2
Any limitation	25.6%	12.8%	28.1%	16.4%	31.5%	25.8%	12.2%	29.6%	17.5%	32.0%
Social limitation	4.3%	1.5%	5.6%	2.3%	5.4%	6.3%	2.8%	7.1%	3.6%	8.2%
Cognition limitation	4.4%	2.2%	6.0%	3.1%	4.7%	6.3%	3.8%	7.9%	4.5%	7.2%
Diabetes	9.4%	7.5%	12.4%	9.4%	8.6%	11.6%	9.5%	14.8%	11.9%	10.4%
Asthma	8.8%	5.3%	10.2%	6.5%	10.0%	9.3%	5.5%	11.8%	7.5%	10.0%
High blood pressure	32.8%	25.7%	43.1%	24.1%	34.4%	34.7%	25.4%	45.0%	26.7%	37.2%
Coronary heart disease	5.6%	2.5%	5.3%	3.7%	7.2%	5.3%	2.7%	4.7%	4.3%	6.6%
Angina	2.7%	1.2%	2.3%	1.8%	3.6%	2.3%	1.3%	1.7%	1.4%	3.3%
Myocardial infarction	3.6%	1.2%	3.6%	1.9%	4.9%	3.8%	1.6%	3.8%	2.4%	5.1%
Stroke	3.6%	1.4%	5.0%	1.9%	4.2%	4.3%	2.1%	6.3%	2.4%	5.1%
Emphysema	2.1%	0.4%	1.6%	0.6%	3.3%	1.9%	0.6%	1.4%	0.6%	3.1%
Cholesterol	30.3%	28.0%	28.7%	24.7%	34.4%	31.6%	28.0%	29.1%	27.0%	36.1%
Arthritis	24.0%	12.5%	27.2%	13.8%	30.0%	26.4%	13.7%	28.0%	16.0%	34.6%
Cancer	8.4%	2.7%	5.0%	3.2%	13.4%	9.5%	2.7%	6.0%	4.5%	15.3%

Continuous variables are presented as *median*

Table 2.2: Median healthcare expenditures stratified by race and characteristics.

		Expenditures in year 2009					Expenditures in year 2016				
Characteristic		Overall	Asians	Blacks	Hispanics	Whites	Overall	Asians	Blacks	Hispanics	Whites
Baseline characteristics											
Age	≤ 45	363	324	284	120	729	389	360	266	181	869
	> 45	2,164	1,149	1,799	921	2,901	2,516	1,817	2,296	1,195	3,399
Male	No	1,326	778	1,110	538	2,236	1,578	1,050	1,274	662	2,700
	Yes	529	349	331	85	1,146	681	486	413	181	1,470
Region	North	1,237	687	964	506	1,924	1,459	723	759	775	2,479
	Midwest	1,173	371	952	305	1,585	1,449	535	1,111	406	1,917
	South	857	342	681	249	1,656	922	497	846	290	2,041
	West	689	633	732	243	1,696	994	1,021	994	418	2,191
SES											
Income	Below poverty	553	376	578	174	1,484	884	1,335	896	386	2,175
	Near poverty	699	342	862	220	1,668	774	251	1,084	280	2,487
	Low	561	477	566	190	1,297	752	589	779	300	1,919
	Middle	818	352	842	276	1,408	922	832	683	377	1,731
	High	1,533	725	1,036	777	2,031	1,692	803	1,225	809	2,339
Education	$<$ High school	494	280	750	210	1,411	696	881	1,003	370	1,908
	High school	840	349	625	262	1,536	956	720	666	300	1,936
	College	1,325	690	1,277	710	1,770	1,533	846	1,265	679	2,098
	Graduate	1,577	883	1,149	652	2,124	1,806	773	1,401	1,129	2,560
Insurance access											
Uninsured	No	1,428	703	1,099	699	2,052	1,445	875	1,121	695	2,292
	Yes	40	40	69	0	150	0	0	0	0	150
Health behaviors											
Smoke	No	985	590	852	289	1,843	1,152	848	918	385	2,252
	Yes	615	240	385	202	1,015	923	332	760	547	1,281
Exercise	No	1,212	490	1,150	300	2,543	1,483	832	1,460	466	2,859
	Yes	757	576	484	259	1,261	857	747	525	320	1,569
Health status											
BMI	< 18.5	617	335	469	170	1,246	1,065	1,028	614	206	2,058
	18.5-24.9	722	497	340	198	1,305	942	733	408	274	1,768
	> 24.9	1,049	642	922	323	1,908	1,233	913	1,043	449	2,307
Mental health	Excellent	613	386	429	151	1,156	644	520	427	229	1,419
	Very good	914	731	594	283	1,573	1,106	735	812	369	1,830
	Good	1,118	605	1,159	346	2,272	1,475	1,234	1,300	503	3,091
	Fair	3,095	2,084	2,808	1,588	4,720	3,410	3,357	3,451	2,216	4,757
	Poor	6,094	1,785	4,290	5,905	7,050	7,108	5,201	7,123	8,329	6,856
Health	Excellent	380	300	184	50	823	409	395	190	115	1,046
	Very good	792	465	513	203	1,436	948	615	559	333	1,689
	Good	1,075	697	1,026	312	2,236	1,441	1,254	1,300	485	3,045
	Fair	2,912	2,044	2,670	1,229	5,614	3,315	2,187	3,386	1,435	6,382
	Poor	8,513	2,756	11,078	6,138	9,785	11,404	7,190	8,147	10,895	13,032
PCS	≤ 50	2,716	1,196	2,199	1,167	4,094	3,574	2,242	3,057	1,890	5,253
	> 50	480	360	328	117	945	549	486	344	196	1,171
MCS	≤ 50	1,251	595	1,178	539	2,211	1,865	1,054	1,836	847	3,039
	> 50	750	499	562	159	1,427	861	659	606	272	1,750
Any limitation	No	539	405	389	171	1,078	619	596	393	250	1,255
	Yes	3,718	2,322	3,138	2,770	4,248	5,237	4,308	4,546	3,774	6,158
Social limitation	No	827	516	648	254	1,536	968	735	739	358	1,839
	Yes	8,852	7,775	8,852	9,997	8,503	9,093	9,005	9,097	9,148	9,140
Cognition limitation	No	833	506	646	250	1,556	980	725	729	353	1,908
	Yes	7,539	4,338	6,407	6,770	8,142	7,977	8,590	7,709	8,196	7,856
Diabetes	No	739	465	518	200	1,429	878	623	593	292	1,751
	Yes	4,745	3,599	5,291	2,693	6,063	5,886	3,142	5,423	3,631	7,624
Asthma	No	828	489	676	244	1,557	992	729	766	346	1,934
	Yes	2,508	1,480	2,092	1,256	3,395	3,115	2,613	2,555	2,207	3,927
High blood pressure	No	469	340	245	127	992	557	460	267	223	1,235
	Yes	2,713	1,896	2,231	1,548	3,639	3,191	2,569	2,662	1,825	4,307
Coronary heart disease	No	800	500	648	250	1,477	995	744	805	357	1,883
	Yes	6,223	4,220	7,982	3,650	6,799	7,394	6,656	7,569	4,526	7,984
Angina	No	863	509	725	263	1,579	1,058	772	857	383	1,973
	Yes	6,129	7,285	6,324	5,600	6,219	8,285	2,465	7,422	7,351	9,445
Myocardial infarction	No	845	515	694	261	1,550	1,022	766	817	372	1,932
	Yes	6,332	4,796	8,095	4,624	6,828	6,937	4,803	6,736	7,116	7,117
Stroke	No	846	507	662	262	1,563	1,017	748	776	372	1,934
	Yes	6,373	4,352	6,307	3,276	7,185	7,268	6,014	7,865	4,446	7,504
Emphysema	No	875	533	732	275	1,586	1,070	777	864	391	1,983
	Yes	6,386	1,665	6,810	6,648	6,599	8,119	903	5,570	7,869	9,330
Cholesterol	No	492	342	366	135	972	570	436	395	212	1,196
	Yes	2,717	1,626	2,686	1,308	3,602	3,295	2,387	3,346	1,722	4,376
Arthritis	No	547	400	383	183	1,028	604	568	435	256	1,208
	Yes	3,622	3,299	2,827	2,468	4,370	4,442	3,827	3,590	3,179	5,076
Cancer	No	757	497	680	250	1,355	916	741	788	358	1,690
	Yes	4,919	5,806	3,713	4,694	5,088	5,697	5,246	5,343	4,072	5,931

Healthcare expenditures are presented as *median*

3 Disparity definition, identification, and estimation

3.1 Path-specific effects as measures of disparity

One approach to measuring racial disparity in healthcare expenditures is to assess whether differences persist if, counterfactually, everyone in the population were assigned to one racial group versus another. Let R denote race and Y denote healthcare expenditures, with counterfactual outcomes $Y(1)$ and $Y(0)$ representing healthcare expenditures if individuals were, hypothetically, members of racial group $R = 1$ (e.g., White) and $R = 0$ (e.g., Black), respectively. The total racial disparity can thus be defined as the population-level contrast $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, which captures the overall effect of race on expenditures.

While this provides a measure of total disparity, interpreting the counterfactuals $Y(1)$ and $Y(0)$ is complicated because race is not a manipulable treatment in the conventional sense, as discussed in the introduction. Rather than representing a direct intervention, race reflects social, historical, and structural factors that shape lived experiences and access to resources [87]. Thus, counterfactual comparisons between racial groups should be understood as quantifying systemic inequities rather than simulating hypothetical experiments in which race itself is altered. Despite these conceptual challenges, the total effect remains a meaningful measure of structural disparities, capturing how racialized differences in social positioning translate into unequal healthcare expenditures. While it does not directly inform intervention strategies, it serves as a diagnostic tool for identifying the magnitude of disparities and motivating further investigation into the mechanisms driving them.

Despite these conceptual challenges, the total effect remains a useful diagnostic tool. However, even if we accept this interpretation, recovering the total effect in

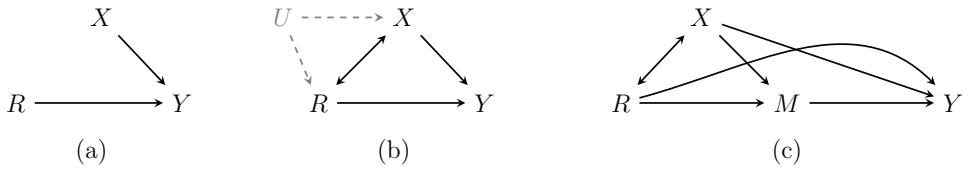


Figure 3.1: A causal diagram where: (a) R affects Y, with X influencing Y but unaffected by R; (b) X and R share a spurious association (bidirected arrow); (c) M mediates the effect of R on Y, with X influencing both M and Y.

observational data presents additional challenges. If race were truly exogenous (Figure 3.1(a)), the observed mean difference, $\mathbb{E}[Y \mid R = 1] - \mathbb{E}[Y \mid R = 0]$, would directly identify the total effect, $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. However, in real-world data, race is often marginally associated with baseline factors such as age, sex, and region. These associations arise possibly because race, as a socially constructed category, is shaped by unmeasured variables—such as parental characteristics, neighborhood context, and historical structural factors—that also influence demographic patterns and geographic distributions [89, 4, 43]. Although these covariates may not be causally related to race (as indicated by bidirected arrows in Figure 3.1(b)), adjusting for them (under common identification assumptions) allows us to recover the total effect via the g-formula [68]: $\int y\{dP(y \mid R = 1, x) - dP(y \mid R = 0, x)\}dP(x)$. The g-formula can be estimated using a regression-based plug-in estimator; however, this approach may yield a biased estimate of the total effect due to outcome model misspecification or improper covariate adjustment, such as blocking part of the racial effect by including downstream mediators like SES or insurance access in the regression model.

Even if we could estimate the total effect without bias, it remains a summary measure that does not reveal how race influences expenditures. To unpack mechanisms driving disparities in the presence of multiple mediators, we employ path-specific effects (PSEs) [70], which decompose the total effect into components corresponding to different mediating pathways. Unlike standard mediation analysis, which typically par-

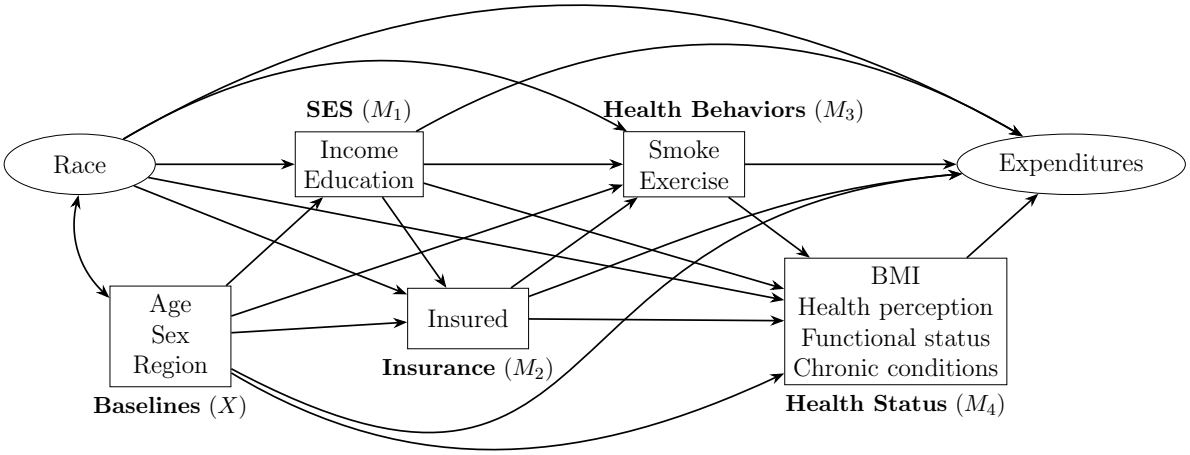


Figure 3.2: A graphical representation of the relations between race, baseline factors, mediating factors, and healthcare expenditures, highlighting pathways via SES, Insurance access, Behavioral factors, and Health Status, as described in Chapter 2.

titions effects into a single direct and indirect pathway (Figure 3.1(c)), PSEs provide a more detailed decomposition by estimating how racial disparities propagate through specific causal pathways. In our context, these pathways correspond to our four key mediators: SES (M_1), insurance access (M_2), health behaviors (M_3), and health status (M_4), as illustrated in Figure 3.2.

We define PSEs as population-level contrasts between counterfactual outcomes under two scenarios. In one, race is set to a reference level ($R = 0$), allowing its influence to propagate naturally through all downstream variables. In the other, along selected pathways, mediators take the values they would have under the non-reference racial group ($R = 1$), while along non-selected pathways, mediators behave as if race were still at the reference level. This follows the path intervention framework of [71] and ensures edge consistency, avoiding the *recanting witness* problem associated with parameter non-identifiability.

We consider five PSEs: the direct effect, corresponding to the direct pathway $\{R \rightarrow$

$Y\}$, and four mediated effects, each capturing the impact of race through a distinct mediator M_k ($k = 1, \ldots, 4$). A mediated effect includes all paths from R to Y passing through M_k, represented as $\{R \to M_k \to Y, R \to M_k \to \ldots \to Y\}$, or more compactly, $\{R \to M_k \rightsquigarrow Y\}$.

To formalize this, let $(r_0, \mathbf{r})$ denote the counterfactual race values along the five specified pathways, where $r_0 \in \{0, 1\}$ and $\mathbf{r} \coloneqq (r_1, r_2, r_3, r_4) \in \{0, 1\}^4$. The setting $\mathbf{r} = \mathbf{0}$ reflects a scenario where all mediators take values under the reference racial group. For a mediated effect through M_k, we set $\mathbf{r}$ to $\mathbf{1}_k$, an indicator vector with the k-th element set to 1, meaning race is set to the non-reference level only along pathways involving $R \to M_k$.

We define the potential outcome:

$$Y(r_0, \mathbf{r}) \coloneqq Y\Big(r_0, \underbrace{M_1(r_1)}_{\coloneqq M_1^c}, \underbrace{M_2\big(r_2, M_1^c\big)}_{\coloneqq M_2^c}, \underbrace{M_3\big(r_3, M_1^c, M_2^c\big)}_{\coloneqq M_3^c}, M_4\big(r_4, M_1^c, M_2^c, M_3^c\big)\Big) \ , \tag{3.1}$$

where mediators are recursively defined as follows: $M_1(r_1)$ (shorthand: M_1^c) is the counterfactual M_1 if $R = r_1$, $M_2(r_2, M_1^c)$ (shorthand: M_2^c) is the counterfactual M_2 if $R = r_2$ and $M_1 = M_1^c$. This recursive structure continues for all four mediators. Using this notation, we define the expected potential outcomes:

$$\gamma_{R \to Y} \coloneqq \mathbb{E}[Y(1, \mathbf{0})] \ , \quad \gamma_{R \to M_k \rightsquigarrow Y} \coloneqq \mathbb{E}[Y(0, \mathbf{1}_k)] \ , \quad \gamma_{\text{ref}} = \mathbb{E}[Y(0, \mathbf{0})] \ . \tag{3.2}$$

The corresponding path-specific effects are defined as:

$$\rho_{R \to Y} \coloneqq \gamma_{R \to Y} - \gamma_{\text{ref}} \ , \qquad \rho_{R \to M_k \rightsquigarrow Y} \coloneqq \gamma_{R \to M_k \rightsquigarrow Y} - \gamma_{\text{ref}} \ . \tag{3.3}$$

In defining the PSEs above, we use a *reference-zero* potential outcome, i.e., $Y(0, \mathbf{0})$. This approach sets race to $R = 1$ (the "active" value) along the pathways of interest while holding it at $R = 0$ (the "inactive" value) elsewhere, and compares the resulting outcome to the baseline $Y(0, \mathbf{0})$. The resulting contrasts are often referred to as *natural path-specific effects* [97]. These estimands reflect the disparity that would remain (or be eliminated) if a single mediator were counterfactually aligned across groups, while

others remained unchanged under the reference race. Importantly, these PSEs are not mutually exclusive and do not decompose the total effect additively. Rather than partitioning the total effect across mediators, we focus on the individual contribution of each pathway in isolation. For comparison, we also consider a sequential decomposition—where the total effect is broken down cumulatively across mediators—in Appendix B [26, 74, 76].

A significant PSE indicates the contribution of specific pathways to population-level racial disparities. Assuming $R = 0$ represents Black population and $R = 1$ represents White population, the effects defined in (3.3), are described in Table 3.1:

Table 3.1: Interpretation of decomposed racial disparity effects

Effect	**Interpretation**
$\rho_{R\to Y}$	Represents structural disparities—the expected difference in healthcare expenditures if individuals were White vs. Black, with all mediators (SES, insurance access, health behaviors, health status) held at levels observed for Black individuals. Often interpreted as the direct effect of perceived race [87], capturing inequities not explained by mediators. Under a weaker interpretation, this is the disparity that persists if mediators for Black individuals are set equal to those of Whites.
$\rho_{R\to M_1\rightsquigarrow Y}$	Captures the effect of race on expenditures through SES. Compares a hypothetical Black population to one where SES takes values had individuals been White, with downstream mediators adapting accordingly. Suggests that addressing socioeconomic barriers could reduce inequities.
$\rho_{R\to M_2\rightsquigarrow Y}$	Captures the effect through insurance access. Compares expenditures for a Black population to one where insurance access reflects that of a White population, with downstream mediators (health behaviors and status) adjusting accordingly. SES remains at Black population levels. Suggests expanded coverage could help reduce inequities.
$\rho_{R\to M_3\rightsquigarrow Y}$	Captures the effect through health behaviors. Compares a Black population to one with White-level health behaviors, with health status adjusting accordingly. SES and insurance access remain at Black population levels. Suggests promoting healthy behaviors may reduce disparities.
$\rho_{R\to M_4\rightsquigarrow Y}$	Captures the effect through health status. Compares expenditures for a Black population to one with White-level health status, holding SES, insurance access, and health behaviors at Black levels. Suggests improving chronic disease management may help reduce disparities.

3.2 Identification assumptions for path-specific effects

Let $\overline{M}_k = (M_1, \cdots, M_k)$ and $\overline{m}_k$ be a realization of $\overline{M}_k$ (for $k = 1, \ldots, 4$), with $\overline{M}_0$ and $\overline{m}_0$ assumed to be the empty sets. We rely on the following assumptions to identify the counterfactual parameters defined in (3.3):

(A1) Consistency, which indicates that observed outcome and mediators match their counterfactuals when race and mediator values are set at observed values; i.e., $Y(r, \overline{m}_4) = Y$ if $R = r$ and $\overline{M}_4 = \overline{m}_4$, and $M_k(r, \overline{m}_{k-1}) = M_k$ if $R = r$ and $\overline{M}_{k-1} = \overline{m}_{k-1}$.

(A2) Positivity, which declares that $P(R = 1 \mid X = x) > 0$ when $P(X = x) > 0$, and $P(R = 1 \mid \overline{M}_k = \overline{m}_k, X = x) > 0$ when $P(\overline{M}_k = \overline{m}_k, X = x) > 0$.

(A3) Ignorability, which states that race is independent of all counterfactuals given X, and any mediator counterfactual is independent of future mediator and outcome counterfactuals given the observed past,

$$Y(r_0, \overline{m}_4), M_4(r_4, \overline{m}_3), M_3(r_3, \overline{m}_2), M_2(r_2, m_1), M_1(r_1) \perp R \mid X \ , \tag{A3.1}$$

$$Y(r_0, \overline{m}_4), M_4(r_4, \overline{m}_3), M_3(r_3, \overline{m}_2), M_2(r_2, m_1) \perp M_1(r) \mid R, X \ , \tag{A3.2}$$

$$Y(r_0, \overline{m}_4), M_4(r_4, \overline{m}_3), M_3(r_3, \overline{m}_2) \perp M_2(r, m_1) \mid M_1, R, X \ , \tag{A3.3}$$

$$Y(r_0, \overline{m}_4), M_4(r_4, \overline{m}_3) \perp M_3(r, \overline{m}_2) \mid \overline{M}_2, R, X \ , \tag{A3.4}$$

$$Y(r_0, \overline{m}_4) \perp M_4(r, \overline{m}_3) \mid \overline{M}_3, R, X \ . \tag{A3.5}$$

Assumptions (A1) and (A2) are standard in the causal inference literature. Assumption (A3) involves "cross-world" independencies, which hold under nonparametric structural equation models with independent errors [65]. In this framework, each variable is generated by an unrestricted structural equation—a nonparametric function of its direct causes (parents in a DAG) and an exogenous error term—where error

terms are assumed to be mutually independent. The cross-world assumptions in (A3) are subject to debate, as they govern interdependencies between race, mediators, and outcomes across hypothetical scenarios that may not co-occur in observable reality. Alternative mediation effect definitions, such as *separable effects* or *stochastic interventions* [29, 75, 56, 28], provide different perspectives on mediation estimands and cross-world identification assumptions. While these approaches offer useful insights, we do not pursue them here.

Under these assumptions, the counterfactual means γ_{ref}, $\gamma_{R\to Y}$, and $\gamma_{R\to M_k \rightsquigarrow Y}$, for $k = 1, 2, 3, 4$, defined in (3.2), can be identified using the *edge g-formula*, as described in [70, 72].

Theorem 3.2.1. *Given Assumptions (A1), (A2), and (A3), the counterfactual means defined in* (3.2)*, are identified as follows:*

$$\gamma_{ref} = \int y dP(y \mid R = 0, x) dP(x) \ ,$$

$$\gamma_{R\to Y} = \int y dP(y \mid \overline{m}_4, R = 1, x) \prod_{k=1}^{4} dP(m_k \mid \overline{m}_{k-1}, R = 0, x) dP(x) \ , \tag{3.4}$$

$$\gamma_{R\to M_k \rightsquigarrow Y} = \int y dP(y \mid \overline{m}_4, R = 0, x) dP(m_k \mid \overline{m}_{k-1}, R = 1, x) \prod_{j=1, j\neq k}^{4} dP(m_j \mid \overline{m}_{j-1}, R = 0, x) dP(x) \ .$$

See a proof in Appendix A.1.

Given the identification functionals in Theorem 3.2.1, the effects defined in (3.3) are simply identified by contrasts of identification functionals for $\gamma_{R\to Y}$ and $\gamma_{R\to M_k \rightsquigarrow Y}$ against γ_{ref}.

There is a substantial body of literature on flexible estimation of causal effects within non/semiparametric models [84, 5, 80, 82, 22]. This includes robust estimation of mediation effects involving one or multiple mediators [79, 57, 11, 97]. In the following section, we develop one-step corrected plug-in estimators using nonparametric influence functions for the identification functionals in (3.4). Our estimators are closely related

to those proposed by [97] for identifiable path-specific effects.

3.3 Estimation techniques and multiply robust estimators

Given n i.i.d. copies of observed data, $\{O_i = (Y_i, \overline{M}_{4,i}, R_i, X_i) : i = 1, \dots n\}$, drawn from distribution P, the effects of interest with the identifying functionals derived from Theorem 3.2.1 can be estimated using plug-in estimates of nuisance functional parameters, including the outcome mean regression and conditional densities of mediators, while empirically evaluating the distribution over covariates X. However, such plug-in estimates (i) may suffer from substantial first-order bias, and (ii) can be computationally challenging due to the need for estimating conditional densities of mixed-type (discrete and continuous) multivariate mediators in our data. In the following, we derive estimators to address these two main limitations. We particularly focus on estimations of counterfactual means $\gamma_{R\to Y}$ and $\gamma_{R\to M_k \rightsquigarrow Y}$, since γ_{ref} is the adjustment functional [68, 65], and the estimation has been extensively discussed in prior literature [5, 80, 83, 84, 22].

To address the *first issue* regarding first-order bias, we can analyze the stochastic properties of the plug-in estimator by utilizing a linear expansion. For an integrable function f defined on the observed data O, let $Pf \coloneqq \int f(o)dP(o)$ denote the expectation under the true distribution P, and let $P_n f \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(O_i)$ represent the empirical average based on the sample. The linear expansion of the plug-in estimator for parameter γ, denoted by $\gamma^{\text{plug-in}}(\hat{Q})$ (where $\hat{Q}$ is the collection of nuisance estimates) is given by: $\gamma^{\text{plug-in}}(\hat{Q}) = \gamma(Q) - P\Phi(\hat{Q}) + R_2(\hat{Q}, Q)$, where Φ denotes the gradient (or influence function) of the parameter, and $R_2(\hat{Q}, Q)$ denotes the remainder terms of second and higher orders from the linear approximation. The term $-P\Phi(\hat{Q})$ is the plug-in's first-order bias, due to substituting $\hat{Q}$ for the true nuisance parameters in $\Phi(Q)$. Although Φ has zero expectation under P (i.e., $P\Phi = 0$), this bias may

still be significant. By deriving the nonparametric influence functions for the counterfactual means, we apply a one-step correction that debiases the plug-in estimator by adjusting for an estimate of its first-order bias (i.e., $-P_n\Phi(\hat{Q})$), yielding the estimator $\gamma^+(\hat{Q}) = \gamma^{\text{plug-in}}(\hat{Q}) + P_n\Phi(\hat{Q})$ [13, 84, 22].

To address the *second issue* regarding density estimation and numerical integration, we parameterize the nonparametric influence functions to bypass these tasks. We rely on the following key nuisance functional components: (i) the propensity score $P(R=1 \mid X)$, denoted as $\pi(X)$; (ii) the binary regressions $P(R=1 \mid \overline{M}_k, X)$ denoted as $g_k(\overline{M}_k, X)$; (iii) the outcome regressions $\mathbb{E}[Y \mid \overline{M}_k, r_0, X]$ denoted as $\mu_k(\overline{M}_k, r_0, X)$; (iv) the sequential regressions $\mathcal{B}_k(\overline{M}_{k-1}, r_k, X) = \mathbb{E}[\mu_k(\overline{M}_k, r_0, X) \mid \overline{M}_{k-1}, r_k, X]$, $\mathcal{C}_{\mathcal{B}_k}(r_1, X) = \mathbb{E}[\mathcal{B}_k(\overline{M}_{k-1}, r_k, X) \mid r_1, X]$, and $\mathcal{C}_{\mu_4}(r_1, X) = \mathbb{E}[\mu_4(\overline{M}_4, r_0, X) \mid r_1, X]$; and (v) the marginal distribution of covariates, P_X. Let $Q = \{\pi, \{g_k, \mu_k, \mathcal{B}_k, \mathcal{C}_{\mathcal{B}_k} : \forall k\}, \mathcal{C}_{\mu_4}\}$ collect all the nuisances. The influence functions for $\gamma_{R\to Y}$ and $\gamma_{R\to M_k \rightsquigarrow Y}$, denoted by $\Phi_{R\to Y}(Q)$ and $\Phi_{R\to M_k\rightsquigarrow Y}(Q)$, respectively, are given as follows: (See detailed derivations in Appendix A.2.)

$$\begin{aligned}\Phi_{R\to Y}(Q)(O_i) = {} & \frac{R_i}{1-\pi(X_i)} \frac{1-g_4(\overline{M}_{4,i}, X_i)}{g_4(\overline{M}_{4,i}, X_i)} \left\{Y_i - \mu_4(\overline{M}_{4,i}, R=1, X_i)\right\} \\ & + \frac{1-R_i}{1-\pi(X_i)} \left\{\mu_4(\overline{M}_{4,i}, R=1, X_i) - \mathcal{C}_{\mu_4}(R=0, X_i)\right\} + \mathcal{C}_{\mu_4}(R=0, X_i) - \gamma_{R\to Y} \ ,\end{aligned} \tag{3.5}$$

$$\begin{aligned}&\Phi_{R\to M_k\rightsquigarrow Y}(Q)(O_i) \\ &= \frac{1-R_i}{1-\pi(X_i)} \frac{g_k(\overline{M}_{k,i}, X_i)}{1-g_k(\overline{M}_{k,i}, X_i)} \frac{1-g_{k-1}(\overline{M}_{k-1,i}, X_i)}{g_{k-1}(\overline{M}_{k-1,i}, X_i)} \left\{Y_i - \mu_k(\overline{M}_{k,i}, R=0, X_i)\right\} \\ &\quad + \frac{R_i}{1-\pi(X_i)} \frac{1-g_{k-1}(\overline{M}_{k-1,i}, X_i)}{g_{k-1}(\overline{M}_{k-1,i}, X_i)} \left\{\mu_k(\overline{M}_{k,i}, R=0, X_i) - \mathcal{B}_k(\overline{M}_{k-1,i}, R=1, X_i)\right\} \\ &\quad + \frac{1-R_i}{1-\pi(X_i)} \left\{\mathcal{B}_k(\overline{M}_{k-1,i}, R=1, X_i) - \mathcal{C}_{\mathcal{B}_k}(r_1, X_i)\right\} + \mathcal{C}_{\mathcal{B}_k}(r_1, X_i) - \gamma_{R\to M_k\rightsquigarrow Y} \ .\end{aligned} \tag{3.6}$$

Given the observed sample, we can use flexible statistical and machine learning models to estimate regressions $\pi, g_k, \mu_k,$ while $\mathcal{B}_k, \mathcal{C}_{\mathcal{B}_k}, \mathcal{C}_{\mu_4}$ can be estimated via a

sequential regression scheme. Estimation of $\mathcal{B}_k$ involves constructing a pseudo-outcome variable $\hat{\mu}_k(\overline{M}_{k,i}, r_0, X_i)$, setting $R_i = r_0$ for all observations. This pseudo-outcome is then regressed on $\overline{M}_{k-1}, X$ using only data points where $R_i = r_k$, yielding estimate $\hat{\mathcal{B}}_k$. Estimation of $\mathcal{C}_{\mathcal{B}_k}$ involves constructing a pseudo-outcome variable $\hat{\mathcal{B}}_k(\overline{M}_{k-1,i}, r_k, X_i)$, setting $R_i = r_k$ for all observations. This pseudo-outcome is then regressed on X using only data points where $R_i = r_1$, yielding estimate $\hat{\mathcal{C}}_{\mathcal{B}_k}$. Finally, $\mathcal{C}_{\mu_4}$ can be estimated via first constructing the a pseudo-outcome variable $\hat{\mu}_4(\overline{M}_{4,i}, r_0, X_i)$, setting $R_i = r_0$ for all observations, and then regressing this pseudo-outcome on X using only data points where $R_i = r_1$, yielding estimate $\hat{\mathcal{C}}_{\mu_4}$. Let $\hat{Q}$ collect the nuisance estimates. Our one-step estimators of $\gamma_{R\to Y}$ and $\gamma_{R\to M_k \rightsquigarrow Y}$, defined in (3.2) and identified in (3.4), are given as follows:

$$\begin{aligned}\gamma^{+}_{R\to Y}(\hat{Q}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ &\frac{R_i}{1-\hat{\pi}(X_i)}\,\frac{1-\hat{g}_4(\overline{M}_{4,i}, X_i)}{\hat{g}_4(\overline{M}_{4,i}, X_i)}\,\big\{Y_i - \hat{\mu}_4(\overline{M}_{4,i}, R=1, X_i)\big\} \\ &+ \frac{1-R_i}{1-\hat{\pi}(X_i)}\big\{\hat{\mu}_4(\overline{M}_{4,i}, R=1, X_i) - \hat{\mathcal{C}}_{\mu_4}(R=0, X_i)\big\} + \hat{\mathcal{C}}_{\mu_4}(R=0, X_i)\Big\}\,,\end{aligned} \tag{3.7}$$

$$\begin{aligned}&\gamma^{+}_{R\to M_k \rightsquigarrow Y}(\hat{Q}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1-R_i}{1-\hat{\pi}(X_i)}\,\frac{\hat{g}_k(\overline{M}_{k,i}, X_i)}{1-\hat{g}_k(\overline{M}_{k,i}, X_i)}\,\frac{1-\hat{g}_{k-1}(\overline{M}_{k-1,i}, X_i)}{\hat{g}_{k-1}(\overline{M}_{k-1,i}, X_i)}\,\big\{Y_i - \hat{\mu}_k(\overline{M}_{k,i}, R=0, X_i)\big\} \\ &\quad + \frac{R_i}{1-\hat{\pi}(X_i)}\,\frac{1-\hat{g}_{k-1}(\overline{M}_{k-1,i}, X_i)}{\hat{g}_{k-1}(\overline{M}_{k-1,i}, X_i)}\big\{\hat{\mu}_k(\overline{M}_{k,i}, R=0, X_i) - \hat{\mathcal{B}}_k(\overline{M}_{k-1,i}, R=1, X_i)\big\} \\ &\quad + \frac{1-R_i}{1-\hat{\pi}(X_i)}\big\{\hat{\mathcal{B}}_k(\overline{M}_{k-1,i}, R=1, X_i) - \hat{\mathcal{C}}_{\mathcal{B}_k}(r_1, X_i)\big\} + \hat{\mathcal{C}}_{\mathcal{B}_k}(r_1, X_i)\Big\}\,.\end{aligned} \tag{3.8}$$

Let $\gamma^{+}(\hat{Q})$ denote either $\gamma^{+}_{R\to Y}(\hat{Q})$ in (3.7) or $\gamma^{+}_{R\to M_k \rightsquigarrow Y}(\hat{Q})$ in (3.8). Asymptotic properties of $\gamma^{+}(\hat{Q})$ can be established through analyzing a linear expansion: $\gamma^{+}(\hat{Q}) - \gamma(Q) = P_n(\Phi(Q)) + (P_n - P)(\Phi(\hat{Q}) - \Phi(Q)) + R_2(\hat{Q}, Q)$. The term $P_n(\Phi(Q))$ is $O_P(n^{-1/2})$ (under central limit theorem), and the term $(P_n - P)(\Phi(\hat{Q}) - \Phi(Q))$ is $o_P(n^{-1/2})$ (under regularity conditions detailed in Appendix A.3). Thus, $\gamma^{+}(\hat{Q})$ is asymptotically linear

if $R_2(\hat{Q}, Q) = o_P(n^{-1/2})$. The following theorem formally states sufficient requirements for the one-step corrected plug-in estimators to be asymptotically linear. Detailed derivations of the remainder terms are provided in Appendix A.3.

Theorem 3.3.1. *Assume the the following $L^2(P)$ convergence rates for the nuisance estimates: $\|\hat{\pi} - \pi\| = o_P(n^{-\frac{1}{a}})$, $\|\hat{g}_k - g_k\| = o_P(n^{-\frac{1}{b_k}})$, $\|\hat{\mathcal{C}}_{\mu_4} - \mathcal{C}_{\mu_4}\| = o_P(n^{-\frac{1}{c}})$, $\|\hat{\mathcal{C}}_{\mathcal{B}_k} - \mathcal{C}_{\mathcal{B}_k}\| = o_P(n^{-\frac{1}{d_k}})$, $\|\hat{\mathcal{B}}_k - \mathcal{B}_k\| = o_P(n^{-\frac{1}{l_k}})$, $\|\hat{\mu}_k - \mu_k\| = o_P(n^{-\frac{1}{m_k}})$ for $k = 1, 2, 3, 4$. Under regularity conditions detailed in Appendix A.3,*

1. *if $\frac{1}{a} + \frac{1}{c} \geq \frac{1}{2}$ and $\frac{1}{b_4} + \frac{1}{m_4} \geq \frac{1}{2}$, then $\sqrt{n}\big(\gamma^+_{R\to Y}(\hat{Q}) - \gamma_{R\to Y}(Q)\big)$ is asymptotically normal with variance equal to $\mathbb{E}[\Phi^2_{R\to Y}(Q)]$;*

2. *if $\frac{1}{a} + \frac{1}{d_k} \geq \frac{1}{2}$, $\frac{1}{b_{k-1}} + \frac{1}{l_k} \geq \frac{1}{2}$ and $\frac{1}{b_k} + \frac{1}{m_k} \geq \frac{1}{2}$, $k = 1, 2, 3, 4$, then $\sqrt{n}\big(\gamma^+_{R\to M_k \rightsquigarrow Y}(\hat{Q}) - \gamma_{R\to M_k \rightsquigarrow Y}(Q)\big)$ is asymptotically normal with variance equal to $\mathbb{E}[\Phi^2_{R\to M_k \rightsquigarrow Y}(Q)]$.*

See a proof in Appendix A.3. Given that $\pi \equiv g_0$, $\mathcal{B}_1 \equiv \mathcal{C}_{\mathcal{B}_1}$, we have $a = b_0$ and $d_1 = l_1$.

The $L^2(P)$ convergence assumptions in Theorem 3.3.1 establish that $R_2(\hat{Q}) = o_P(n^{-1/2})$, even when flexible models with slower convergence rates than $n^{-1/2}$ are used for nuisance functional estimations. Moreover, Theorem 3.3.1 implies certain robustness behaviors for consistency of $\gamma^+(\hat{Q})$, formalized in the following corollary. (See a proof in Appendix A.3.)

Corollary 3.3.2. *Under regularity conditions detailed in Appendix A.3, the one-step estimators are consistent if at least one of the following estimates are consistent:*

1. *For $\gamma^+_{R\to Y}(\hat{Q})$: (i) $\hat{\pi}$ and $\hat{g}_4$, (ii) $\hat{\pi}$ and $\hat{\mu}_4$, or (iii) $\hat{\mathcal{C}}_{\mu_4}$ and $\hat{\mu}_4$;*

2. *For $\gamma^+_{R\to M_k \rightsquigarrow Y}(\hat{Q})$, $k = 1, 2, 3, 4$: (i) $\hat{\pi}$, $\hat{g}_{k-1}$, and $\hat{g}_k$, (ii) $\hat{\pi}$, $\hat{g}_{k-1}$, and $\hat{\mu}_k$, (iii) $\hat{\pi}$, $\hat{\mathcal{B}}_k$, and $\hat{\mu}_k$, or (iv) $\hat{\mathcal{C}}_{\mathcal{B}_k}$, $\hat{\mathcal{B}}_k$, and $\hat{\mu}_k$.*

Given that $\pi \equiv g_0$ and $\mathcal{B}_1 \equiv \mathcal{C}_{\mathcal{B}_1}$, when $k = 1$, the third set of nuisance estimates for consistency of $\gamma^+_{R \to M_k \rightsquigarrow Y}(\hat{Q})$ is a superset of the fourth condition, making it redundant. Corollary 3.3.2 suggests that $\gamma^+(\hat{Q})$ can achieve consistency even if certain parts of the underlying observed joint distribution are misspecified.

One-step corrected plug-in estimates of PSEs $\rho_{R \to Y}$ and $\rho_{R \to M_k \rightsquigarrow Y}$, defined in (3.3), can be obtained via one-step corrected plug-in estimates of $\gamma_{R \to Y}$, $\gamma_{R \to M_k \rightsquigarrow Y}$, and γ_{ref}. Such an estimator for γ_{ref} is known as the *augmented inverse probability weighted* estimator, which we denote by $\gamma^+_{\text{ref}}(\hat{Q})$, where $\hat{Q}$ is a slight abuse of notation that refers to estimates of the propensity score $P(R = 1 \mid x)$, noted as $\pi(x)$, and the outcome regressions $\mathbb{E}[Y \mid r, x]$, represented as $\mu_0(r, x)$. Thus, we can write:

$$\rho^+_{R \to Y}(\hat{Q}) = \gamma^+_{R \to Y}(\hat{Q}) - \gamma^+_{\text{ref}}(\hat{Q}) \ , \quad \rho^+_{R \to M_k \rightsquigarrow Y}(\hat{Q}) = \gamma^+_{R \to M_k \rightsquigarrow Y}(\hat{Q}) - \gamma^+_{\text{ref}}(\hat{Q}) \ . \tag{3.9}$$

4 Empirical analysis of the MEPS data

We apply our methodological framework to the MEPS data, described in Chapter 2.

4.1 Implementation details

To estimate the PSEs of interest using the estimators outlined in (3.7), (3.8), and (3.9), we fit each nuisance function-valued parameter in $Q = \{\pi, \{g_k, \mu_k, \mathcal{B}_k, \mathcal{C}_{\mathcal{B}_k} : \forall k\}, \mathcal{C}_{\mu_4}\}$, as described in Chapter 3.3, using super learners. This ensemble learning method combines flexible statistical and machine learning models via cross-validation to mitigate model misspecification and improve predictive accuracy [83, 67]. We include `mean`, `glm`, `glm.interaction`, `gam`, `glmnet`, `earth`, `ksvm`, `xgboost`, `randomForest`, `dbarts` as candidate learners.

When estimating outcome regressions $\mu_k(\overline{M}_k, r_0, X)$ using MEPS data, challenges arise from zero-inflated and right-skewed distribution of healthcare expenditures, as shown in Figure 4.1. In health economics, the two-part model is widely used to address such complexities [52, 1, 73]. This approach models the data as a mixture by separating it into two parts [10]: the first part estimates the probability of a non-zero response $P(Y > 0 \mid \overline{M}_k, R, X)$, and the second part models the distribution of the positive responses, $P(Y \mid Y > 0, \overline{M}_k, R, X)$. The conditional mean of the outcome can then be expressed as $\mathbb{E}[Y \mid \overline{M}_k, R, X] = P(Y > 0 \mid \overline{M}_k, R, X) \times \mathbb{E}[Y \mid Y > 0, \overline{M}_k, R, X]$. The probability of a non-zero response can be readily estimated using flexible learners, while the mean of positive responses is often modeled with generalized linear models (GLMs) that assume a Gamma or Lognormal error distribution to handle right-skewed data [14, 54]. Wu et al. [94] propose an alternative two-stage super learner, which includes GLMs with Gamma distribution and various link functions as candidate learners. Using these estimation methods, predictions for the i-th observation is obtained by combining

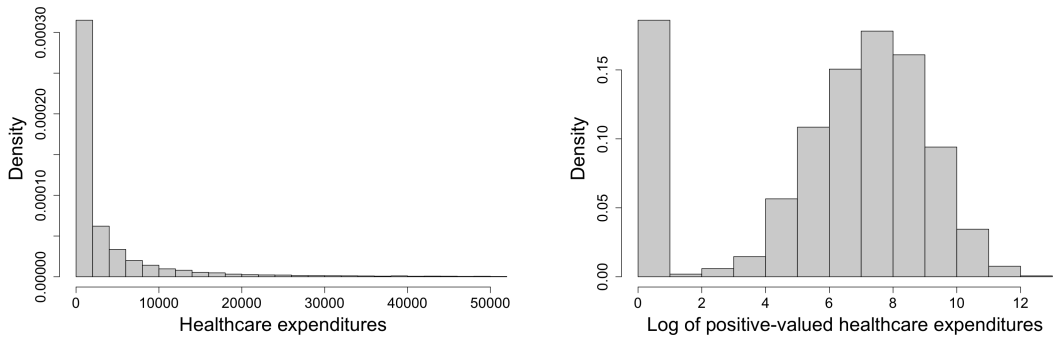


Figure 4.1: The empirical distribution of healthcare expenditures in the MEPS data.

the results from the two-part model, i.e., $\hat{\mu}_k(\overline{M}_{k,i}, r_0, X_i) = \hat{P}(Y > 0 \mid \overline{M}_{k,i}, r_0, X_i) \times \hat{\mathbb{E}}[Y \mid Y > 0, \overline{M}_{k,i}, r_0, X_i]$.

We adapted a two-part model to estimate the regressions $\mu_k(\overline{M}_k, r_0, X)$, incorporating a log-transformation of the positive healthcare expenditures. Given the skewed distribution of these expenditures, reporting effects on the arithmetic mean scale (i.e., without log-transformation) can be overly sensitive to extreme values. By applying a log-transformation, we instead report the effects on the geometric mean scale, which is less influenced by extremes and thus more appropriate for skewed data, as detailed below [6].

Assuming healthcare expenditures are positive, we consider the log-transformation of the potential outcomes defined in (3.1), and express the direct and indirect effects as:

$$\exp(\rho_{R\to Y}) = \exp(\mathbb{E}[\log Y(1,\mathbf{0})] - \mathbb{E}[\log Y(0,\mathbf{0})]) \approx G_n(Y(1,\mathbf{0}))/G_n(Y(0,\mathbf{0}))\ ,$$

$$\exp(\rho_{R\to M_k \rightsquigarrow Y}) = \exp(\mathbb{E}[\log Y(0,\mathbf{1}_k)] - \mathbb{E}[\log Y(0,\mathbf{0})]) \approx G_n(Y(0,\mathbf{1}_k))/G_n(Y(0,\mathbf{0}))\ ,$$

where $G_n(f)$ denotes the geometric mean of f, i.e., $G_n(f) = \{\prod_{i=1}^n f_i\}^{1/n}$. These expressions represent the ratio of geometric means of the respective potential outcomes, providing a clear interpretation of the relative difference in healthcare expenditures

attributable to racial differences. A value greater than one suggests higher geometric mean expenditures under the active condition, while a value less than one indicates lower expenditures. Identification and estimation, as discussed in Chapther 3, extend naturally by considering the log-transformed positive healthcare expenditures as the observed outcome. The delta method is then used to compute the variance of the transformation for inference. The calculated effects are then reported using exponential re-transformation, placing them to the original scale with a geometric mean interpretation.

To address the zero-inflated nature of expenditures in our data, we redefine the observed outcome as $\mathbb{I}(Y > 0) \times \log Y$, ensuring that the log transformation is restricted to positive responses. We then modify the second part of the two-part model to estimate $\mathbb{E}[\log Y \mid Y > 0, \overline{M}_k, r_0, X]$ under the assumption of a normal error distribution. Effects are reported as $\exp(\rho_{R\to Y})$ and $\exp(\rho_{R\to M_k \rightsquigarrow Y})$, interpreted as the ratio of geometric means of positive potential outcomes, adjusted for the probability of observing positive expenditures. This approach accounts for the zero-inflated nature of the data while maintaining the geometric mean interpretation. Further details are provided in Appendix C.1.

4.2 Empirical results

The total effect and natural PSEs are reported as ratios of scaled geometric means in Table 4.1. Meanwhile, we provide cumulative PSEs using sequential decomposition in Appendix B.1.

The total effect was significant across all six racial group comparisons (White vs. Black, White vs. Asian, White vs. Hispanic, Black vs. Asian, Black vs. Hispanic, and Asian vs. Hispanic) in 2009. This effect reflects how expenditures for one racial group would change on average if, hypothetically, they belonged to another racial group.

Whites consistently exhibited the highest scaled geometric mean expenditures in comparisons involving other racial groups. One potential explanation for this pattern is systemic advantages in healthcare access and utilization, as suggested in [52, 3, 32]. Comparisons among minority groups revealed additional disparities: for instance, Hispanics consistently had the lowest counterfactual expenditures, further highlighting structural inequities across racial hierarchies. These racial disparities persisted in 2016, except for the non-significant total effect between Black and Asian groups. The total effect for the White vs. Black comparison increased in 2016, consistent with findings by Dickman et al. [31], who reported a widening gap in total healthcare expenditures between White and Black populations from the periods 2010–2013 to 2014–2019. Conversely, the total effects for the White vs. Asian, White vs. Hispanic, Black vs. Asian, and Black vs. Hispanic comparisons declined in 2016, suggesting a partial narrowing of disparities among these groups. These patterns may reflect changes in socioeconomic conditions, policy environments, or healthcare access across racial groups, though further research is needed to fully understand these trends.

The effect through SES (M_1), assessed by income and education, was significant across five racial group comparisons, except for White vs. Asian, in both 2009 and 2016. This effect can be interpreted as aligning the SES distribution (conditioned on covariates) across the two racial groups. In 2009, if a hypothetical Black or Hispanic population had an SES distribution aligned with that of Whites, the scaled geometric mean expenditures would increase to 1.114 (95% CI: 1.054–1.173) times or 1.450 (95% CI: 1.344–1.557) times, respectively. Similarly, if a hypothetical Asian or Hispanic population had an SES distribution aligned with that of Blacks, the scaled geometric mean expenditures would decrease by 16.5% or increase by 19.2%, respectively. Notably, aligning the SES of a hypothetical Hispanic population with that of Asians would result in a nearly doubling of scaled geometric mean expenditures. These results sug-

gest that SES plays a major role in racial disparities in healthcare expenditures. Asians tend to have high SES levels [93], while Black and Hispanic populations face higher poverty rates and lower levels of higher education compared to Whites and Asians [19]. These SES variations help explain some of the disparities observed in SES-mediated effects. In 2016, SES-mediated effects showed a slight increase compared to 2009, suggesting a growing role of income and education disparities in shaping healthcare expenditures. These findings highlight SES as a key driver of racial disparities, both through direct economic effects on healthcare access and through its influence on other mediators, including insurance access, health behaviors, and health status.

The effect through health insurance (M_2) was significant across all racial group comparisons, except for White vs. Black, in 2009. This effect can be interpreted as the impact of an alignment of insurance access distribution (conditioned on covariates and SES) between racial groups. If the insurance coverage of a hypothetical Asian population was aligned with that of Whites or Blacks, the scaled geometric mean expenditures would increase by 9.1% or 7.9%, respectively. Similarly, if the insurance coverage of a hypothetical Hispanic population was aligned with that of Whites, Blacks, or Asians, the scaled geometric mean expenditures would increase to 1.372 (95% CI: 1.306–1.439), 1.478 (95% CI: 1.393–1.562), or 1.265 (95% CI: 1.176–1.355) times, respectively. Notably, Hispanics had the highest rate of being uninsured in 2009—more than three times that of Whites. By 2016, the insurance-mediated disparities had disappeared in the White vs. Asian and Black vs. Asian comparisons, coinciding with a decline in observed uninsured rates across all racial groups, and particularly small differences between Asians and Whites. One contributing factor could be the Affordable Care Act, enacted in 2010 and fully implemented in 2014, which expanded coverage for economically disadvantaged minorities [35, 63, 16]. However, significant insurance-mediated disparities persisted in all racial group comparisons involving Hispanics. In fact, the

disparities increased in the White vs. Hispanic and Asian vs. Hispanic comparisons (1.380, 95% CI: 1.318–1.441 and 1.320, 95% CI: 1.245–1.395, respectively). Although overall insurance coverage improved, Hispanics continued to have the highest rate of uninsurance, and the gap in healthcare expenditures between insured and uninsured groups widened in 2016, underscoring the growing importance of insurance in healthcare disparities. Barriers for Hispanics may include unclear eligibility policies, difficulty navigating enrollment processes, and language or literacy challenges [40, 91]. Without insurance, individuals often delay seeking care, while having coverage facilitates access and may increase overall expenditures through more timely care [34].

Although small, the effect through health behaviors (M_3), assessed by smoking status and physical activity, was significant only in the White vs. Hispanic comparison in 2009 (1.076, 95% CI: 1.014–1.137), and in the Asian vs. Hispanic comparison in 2016 (1.022, 95% CI: 1.006–1.038). Consistent with the observed data, smoking prevalence was higher among Whites—nearly twice that of Hispanics. Given that smoking is strongly associated with an elevated risk of diseases such as cancer, respiratory, and cardiovascular conditions [60], it contributes substantially to the overall healthcare costs [1]. In contrast, the proportion of individuals who regularly exercised was marginally lowest among Asians in 2016. Exercise plays a critical role in improving health at both individual and population levels [42]. Overall, these findings underscore the influence of health behaviors on healthcare expenditure disparities, highlighting both the risks associated with smoking and the opportunities for intervention through increased physical activity and preventive care.

Health status (M_4) emerged as an important mediator in healthcare expenditure disparities. Prior studies have shown that, compared with Whites, minorities tend to report poorer self-rated health and are more likely to suffer from chronic conditions due to lower SES, limited insurance access, and less favorable living environments

[12, 48, 91]. These factors would typically suggest that minorities bear higher medical spending burdens relative to Whites [21]. However, when focusing solely on the effects mediated through health status—excluding the influence of SES, insurance, and health behaviors—our study revealed a different pattern. In 2009, health status-mediated effect was significant for all racial group comparisons except for the White vs. Black comparison, and by 2016, this effect was significant across all racial group comparisons. Specifically, if the health status of a hypothetical Black, Asian, or Hispanic population were aligned with that of Whites, their geometric mean expenditures would increase by 10.1%, 43.7%, and 53.8%, respectively in 2016. Likewise, aligning the health status of a hypothetical Asian or Hispanic population with that of Blacks would increase expenditures by factors of 1.393 and 1.253, respectively, whereas aligning the health status of a hypothetical Hispanic population with that of Asians would reduce expenditures to 79.6%. The divergence between our findings and previous literature may be attributable to a higher observed disease prevalence among Whites, which could reflect both their greater access to screening and diagnostic services [33] and potential differences in genetic, dietary, or other inherent factors.

The direct effect of race was only significant in comparisons between Whites and any minority group in 2009, and not significant between any two minority groups. One explanation for this direct effect is that some factors were not included in the mediation analysis, leading to the direct effect capturing the influence of unobserved mediators. For instance, early life adversity—such as poverty, abuse, and traumatic stress, which vary by race and SES—has been shown to affect multiple indicators of physical and mental health later in life, ultimately influencing healthcare expenditures [69]. Another potential explanation is structural racism. A systematic review has demonstrated that healthcare professionals' implicit biases are associated with treatment decisions, adherence to treatment recommendations, and patient health outcomes [38]. These biases

may result in poorer communication during medical visits and lower ratings of care, leading minority patients to be less willing to adhere to medical advice [51, 25]. By 2016, the direct effect in the White vs. Asian and White vs. Hispanic comparisons declined, while those for Whites vs. Blacks, Blacks vs. Hispanics, and Asians vs. Hispanics deviated significantly from 1, suggesting an increase in disparities in healthcare expenditures attributable to race for these groups. This shift underscores persistent and evolving structural inequities and points to systemic biases that disproportionately affect certain racial groups.

In summary, our analysis reveals persistent racial disparities in healthcare expenditures, with Whites generally exhibiting higher expenditures compared to minority groups. In 2009, significant disparities emerged across all racial comparisons, primarily mediated by differences in SES and health status, while insurance coverage also played a critical role—particularly in differentiating outcomes for Hispanics. By 2016, although some insurance-mediated gaps (notably for Asians) narrowed, significant disparities persisted, especially for Hispanics, underscoring that insurance remains a key factor alongside SES and health status. These findings highlight the multifaceted drivers of healthcare inequities and underscore the need for targeted interventions—such as enhancing educational opportunities for minority populations, expanding accessible insurance coverage, and equipping healthcare providers with training to recognize and address implicit biases—to mitigate these disparities, while also encouraging further research to explore additional pathways and unmeasured factors contributing to these outcomes.

Table 4.1: Natural path-specific effects across racial group comparisons (scaled geometric mean ratios)

	MEPS data in year 2009			**MEPS data in year 2016**		
Path	Effect	95% CI	p-value	Effect	95% CI	p-value
Whites vs Blacks*						
$R \to M1 \rightsquigarrow Y$	1.114	1.054 — 1.173	**<0.001**	1.191	1.124 — 1.259	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.017	0.984 — 1.050	0.321	1.005	0.977 — 1.033	0.704
$R \to M3 \rightsquigarrow Y$	0.981	0.959 — 1.003	0.089	1.013	0.992 — 1.035	0.219
$R \to M4 \to Y$	1.023	0.954 — 1.092	0.513	1.101	1.023 — 1.179	**0.011**
$R \to Y$	1.772	1.616 — 1.929	**<0.001**	1.869	1.688 — 2.050	**<0.001**
Total effect	2.138	1.894 — 2.382	**<0.001**	2.390	2.108 — 2.672	**<0.001**
Whites vs Asians*						
$R \to M1 \rightsquigarrow Y$	0.975	0.884 — 1.067	0.598	0.935	0.812 — 1.058	0.299
$R \to M2 \rightsquigarrow Y$	1.091	1.024 — 1.157	**0.007**	1.023	0.990 — 1.056	0.175
$R \to M3 \rightsquigarrow Y$	0.970	0.903 — 1.036	0.373	0.975	0.931 — 1.019	0.269
$R \to M4 \to Y$	1.418	1.242 — 1.594	**<0.001**	1.437	1.247 — 1.626	**<0.001**
$R \to Y$	2.399	2.073 — 2.724	**<0.001**	1.944	1.655 — 2.233	**<0.001**
Total effect	2.863	2.377 — 3.350	**<0.001**	2.446	2.033 — 2.859	**<0.001**
Whites vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.450	1.344 — 1.557	**<0.001**	1.537	1.423 — 1.652	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.372	1.306 — 1.439	**<0.001**	1.380	1.318 — 1.441	**<0.001**
$R \to M3 \rightsquigarrow Y$	1.076	1.014 — 1.137	**0.016**	1.047	0.996 — 1.099	0.073
$R \to M4 \to Y$	1.426	1.322 — 1.531	**<0.001**	1.538	1.419 — 1.656	**<0.001**
$R \to Y$	2.097	1.916 — 2.279	**<0.001**	1.938	1.767 — 2.109	**<0.001**
Total effect	4.634	4.141 — 5.128	**<0.001**	4.297	3.823 — 4.771	**<0.001**
Blacks vs Asians*						
$R \to M1 \rightsquigarrow Y$	0.835	0.721 — 0.949	**0.004**	0.820	0.710 — 0.929	**0.001**
$R \to M2 \rightsquigarrow Y$	1.079	1.009 — 1.149	**0.027**	1.024	0.976 — 1.072	0.325
$R \to M3 \rightsquigarrow Y$	0.974	0.931 — 1.017	0.233	0.996	0.956 — 1.037	0.856
$R \to M4 \to Y$	1.440	1.242 — 1.637	**<0.001**	1.393	1.213 — 1.573	**<0.001**
$R \to Y$	1.044	0.876 — 1.212	0.610	0.882	0.744 — 1.019	0.092
Total effect	1.307	1.032 — 1.583	**0.029**	0.979	0.782 — 1.175	0.831
Blacks vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.192	1.130 — 1.254	**<0.001**	1.184	1.126 — 1.241	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.478	1.393 — 1.562	**<0.001**	1.405	1.336 — 1.474	**<0.001**
$R \to M3 \rightsquigarrow Y$	1.023	0.986 — 1.060	0.225	1.023	0.976 — 1.069	0.337
$R \to M4 \to Y$	1.302	1.202 — 1.402	**<0.001**	1.253	1.161 — 1.344	**<0.001**
$R \to Y$	1.024	0.943 — 1.104	0.568	0.879	0.802 — 0.956	**0.002**
Total effect	2.085	1.774 — 2.396	**<0.001**	1.698	1.454 — 1.941	**<0.001**
Asians vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.768	1.569 — 1.967	**<0.001**	1.904	1.701 — 2.106	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.265	1.176 — 1.355	**<0.001**	1.320	1.245 — 1.395	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.999	0.980 — 1.017	0.891	1.022	1.006 — 1.038	**0.006**
$R \to M4 \to Y$	0.788	0.719 — 0.857	**<0.001**	0.796	0.706 — 0.885	**<0.001**
$R \to Y$	1.015	0.939 — 1.091	0.697	1.164	1.069 — 1.259	**0.001**
Total effect	1.855	1.521 — 2.189	**<0.001**	1.866	1.540 — 2.192	**<0.001**

*Reference group; M_1: SES, M_2: Insurance, M_3: Health behaviors, M_4: Health status.

5 Simulation studies

In this chapter, we evaluate our one-step estimators for PSEs using two simulation studies. The first study demonstrates theoretical properties and the robustness of the proposed estimators and illustrates benefits of flexible nuisance parameter estimation by highlighting the poor performance of estimators based on misspecified parametric models. The second assesses the finite sample performance with data mimicking our real-data application.

5.1 Simulation 1: Asymptotic properties and robustness

This simulation evaluates the estimators' asymptotic properties and robustness to model misspecification for Corollary 3.3.2. Data are generated using four uniform covariates, a binary treatment, four ordered univariate continuous mediators (normally distributed), and a normally distributed outcome $(X_1, X_2, X_3, X_4, R, M_1, M_2, M_3, M_4, Y)$. This simulation runs for sample sizes of 250, 500, 1000, 2000, 4000, and 8000, with 1000 replications per scenario.

$$
\begin{aligned}
X_1, X_2, X_3, X_4 &\overset{iid}{\sim} \text{Uniform}(0,1),\\
R &\sim \text{Bernoulli}(\text{expit}(V_R\begin{bmatrix}1 & X\end{bmatrix}^T),\\
M_1 &\sim \mathcal{N}(V_{M_1}\begin{bmatrix}1 & X & R\end{bmatrix}^T, 1),\\
M_2 &\sim \mathcal{N}(V_{M_2}\begin{bmatrix}1 & X & R & M_1\end{bmatrix}^T, 1),\\
M_3 &\sim \mathcal{N}(V_{M_3}\begin{bmatrix}1 & X & R & M_1 & M_2\end{bmatrix}^T, 1),\\
M_4 &\sim \mathcal{N}(V_{M_4}\begin{bmatrix}1 & X & R & M_1 & M_2 & M_3\end{bmatrix}^T, 1),\\
Y &\sim \mathcal{N}(V_Y\begin{bmatrix}1 & X & R & M_1 & M_2 & M_3 & M_4\end{bmatrix}^T, 1).
\end{aligned} \tag{5.1}
$$

Specifically, the coefficients are:

$$V_R = (-0.10, 1.00, 0.20, -0.40, 0.80),$$
$$V_{M_1} = (-0.13, 0.23, -0.18, 0.15, -0.16, 0.13),$$
$$V_{M_2} = (-0.11, -0.06, 0.20, 0.25, 0.02, -0.12, 0.16),$$
$$V_{M_3} = (-0.24, -0.08, -0.15, 0.03, 0.14, 0.06, -0.14, 0.09),$$
$$V_{M_4} = (-0.13, -0.09, -0.04, 0.10, -0.25, -0.05, -0.08, 0.19, -0.20),$$
$$V_Y = (0.43, 0.29, 0.28, -0.26, -0.38, 0.18, 0.39, -0.22, -0.13, 0.28).$$

First, we examine the asymptotic properties of the estimators by evaluating the convergence of the root-n-scaled bias and n-scaled variance. The proposed one-step estimators for for counterfactual means ($\gamma^+_{R\to Y}$ and $\gamma^+_{R\to M_k \rightsquigarrow Y}$) are constructed using estimates of nuisance functions $Q = \{\pi, \{g_k, \mu_k, \mathcal{B}_k, \mathcal{C}_{\mathcal{B}_k} : \forall k\}, \mathcal{C}\mu_4\}$. These nuisance functions can be consistently estimated via GLMs based on linear combinations of the predictors [97]:

$$\begin{aligned} \pi = \text{expit}(\theta_0 \begin{bmatrix} 1 & X \end{bmatrix}^T), \quad g_k = \text{expit}(\theta_k \begin{bmatrix} 1 & X & \overline{M}_k \end{bmatrix}^T)), \\ \mu_k = \alpha_k \begin{bmatrix} 1 & X & R & \overline{M}_k \end{bmatrix}^T, \quad \mathcal{B}_k = \delta_{k-1} \begin{bmatrix} 1 & X & R & \overline{M}_{k-1} \end{bmatrix}^T, \\ \mathcal{C}_{\mathcal{B}_k} = \nu_{\mathcal{B}_k} \begin{bmatrix} 1 & X & R \end{bmatrix}^T, \quad \mathcal{C}_{\mu_4} = \nu_{\mu_4} \begin{bmatrix} 1 & X & R \end{bmatrix}^T. \end{aligned} \tag{5.2}$$

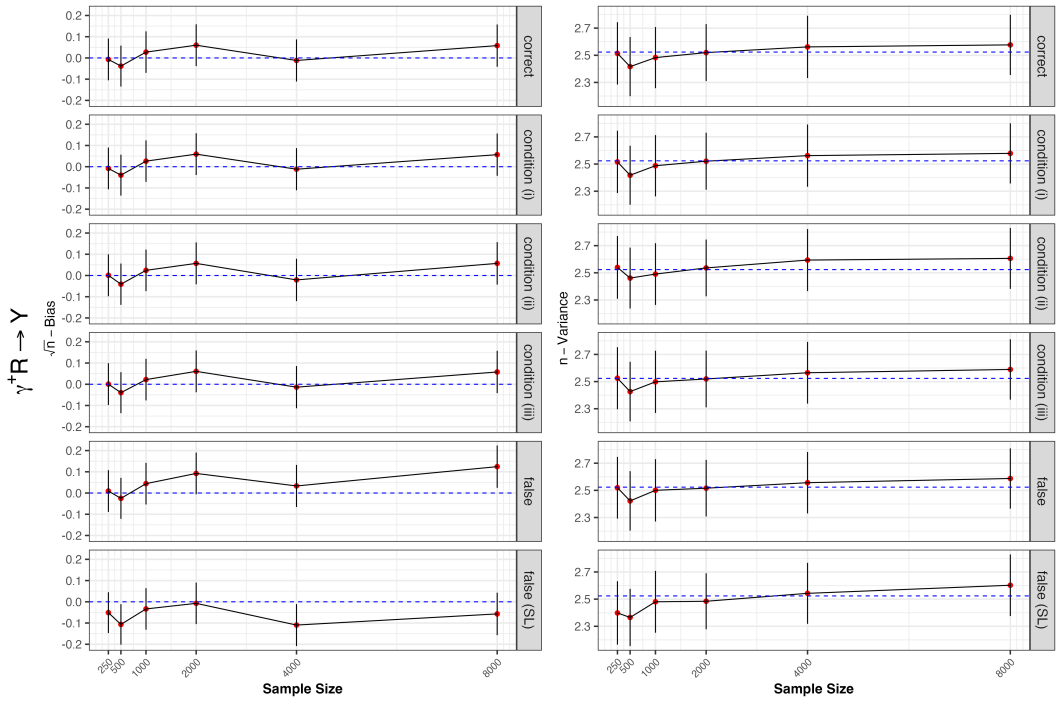
Then, we evaluate the consistency of $\hat{\gamma}^+_{R\to Y}$ under three conditions: (i) only $\hat{\pi}$ and $\hat{g}_4$ are consistent; (ii) only $\hat{\pi}$ and $\hat{\mu}_4$ are consistent; (iii) only $\hat{\mathcal{C}}_{\mu_4}$ and $\hat{\mu}_4$ are consistent. Similarly, the the consistency of $\hat{\gamma}^+_{R\to M_1 \rightsquigarrow Y}$ is evaluated under three conditions:(i) only $\hat{\pi}$ and $\hat{g}_1$ are consistent; (ii) only $\hat{\pi}$, and $\hat{\mu}_1$ are consistent; (iii) only $\hat{\mathcal{B}}_1$, and $\hat{\mu}_1$ are consistent. For $k = 2, 3, 4$, the the consistency of $\hat{\gamma}^+_{R\to M_k \rightsquigarrow Y}(\hat{Q})$, $k = 2, 3, 4$ is evaluated under four conditions: (i) only $\hat{\pi}$, $\hat{g}_{k-1}$, and $\hat{g}_k$ are consistent; (ii) only $\hat{\pi}$, $\hat{g}_{k-1}$ and $\hat{\mu}_k$

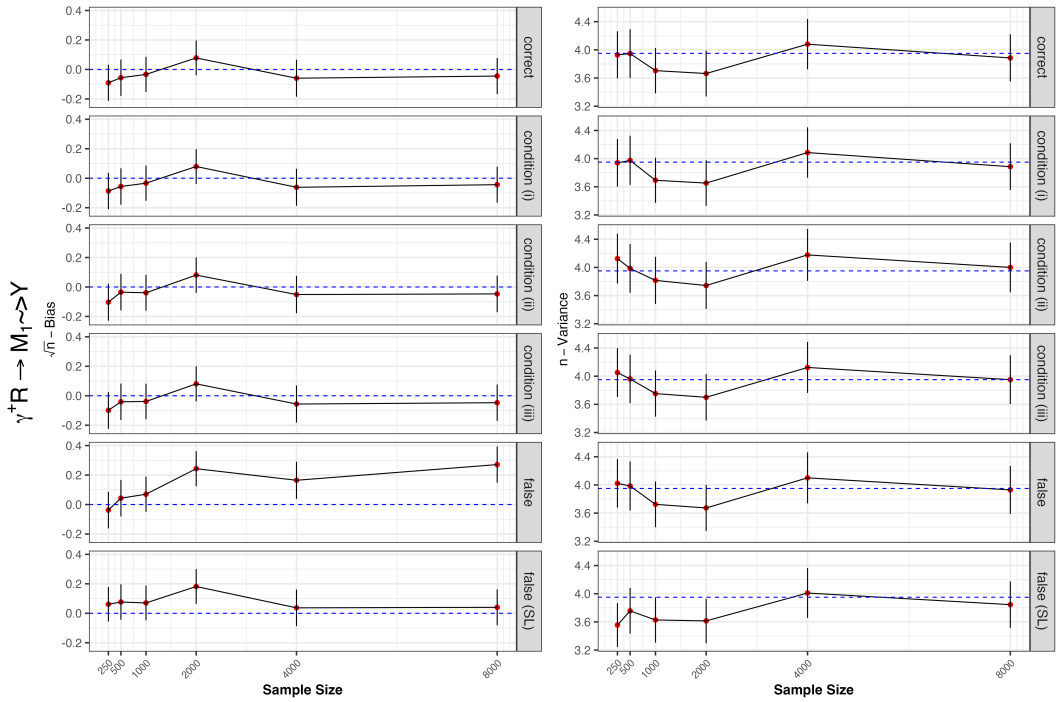
are consistent; (iii) only $\hat{\pi}$, $\hat{\mathcal{B}}_k$ and $\hat{\mu}_k$ are consistent; (iv) only $\hat{\mathcal{C}}_{\mathcal{B}_k}$, $\hat{\mathcal{B}}_k$, and $\hat{\mu}_k$ are consistent. We obtain misspecified nuisance estimates by applying nonlinear transformations to the covariates. Specifically, a set of false covariates is generated from the correct covariates X as $X^{\text{false}} = (X_1^2, e^{X_2}, (X_3)^{0.3}, (X_4 + (X_3)^{0.3})/(e^{X_2} + X_1^2))$, which are then used to construct misspecified functions for Q^{false} via GLMs:

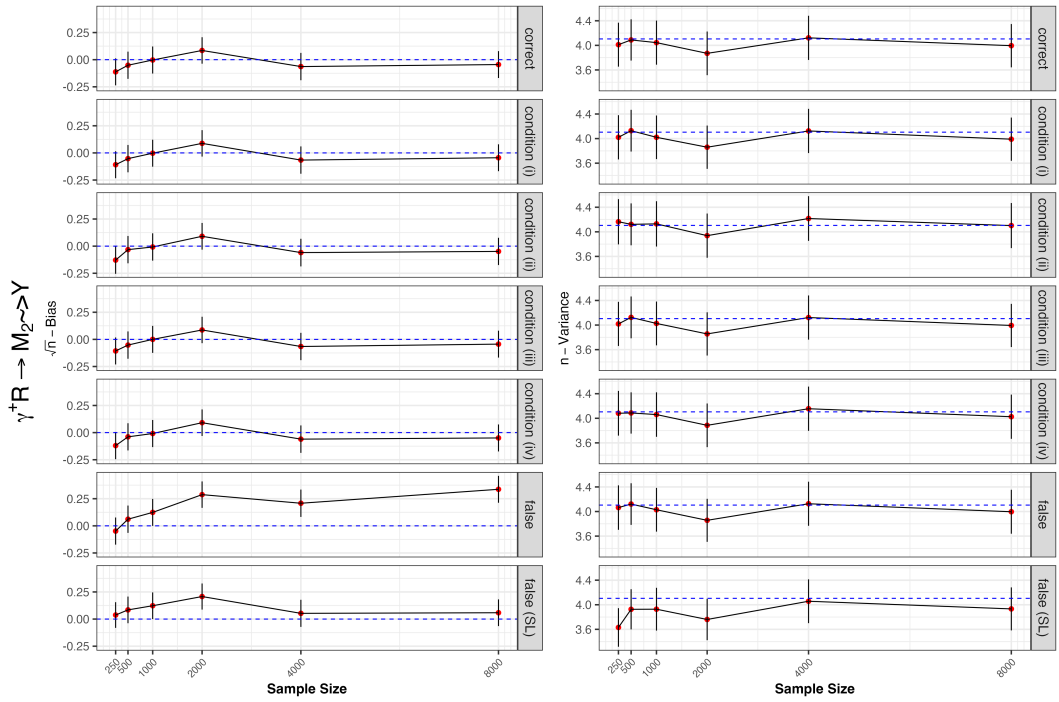
$$\begin{aligned}
\pi^{\text{false}} &= \text{expit}(\theta_0^* \begin{bmatrix} 1 & X^{\text{false}} \end{bmatrix}^T), \quad g_k^{\text{false}} = \text{expit}(\theta_k^* \begin{bmatrix} 1 & X^{\text{false}} & \overline{M}_k \end{bmatrix}^T)), \\
\mu_k^{\text{false}} &= \alpha_k^* \begin{bmatrix} 1 & X^{\text{false}} & R & \overline{M}_k \end{bmatrix}^T, \quad \mathcal{B}_k^{\text{false}} = \delta_{k-1}^* \begin{bmatrix} 1 & X^{\text{false}} & R & \overline{M}_{k-1} \end{bmatrix}^T, \\
\mathcal{C}_{\mathcal{B}_k}^{\text{false}} &= \nu_{\mathcal{B}_k}^* \begin{bmatrix} 1 & X^{\text{false}} & R \end{bmatrix}^T, \quad \mathcal{C}_{\mu_4}^{\text{false}} = \nu_{\mu_4}^* \begin{bmatrix} 1 & X^{\text{false}} & R \end{bmatrix}^T. \qquad (5.3)
\end{aligned}$$

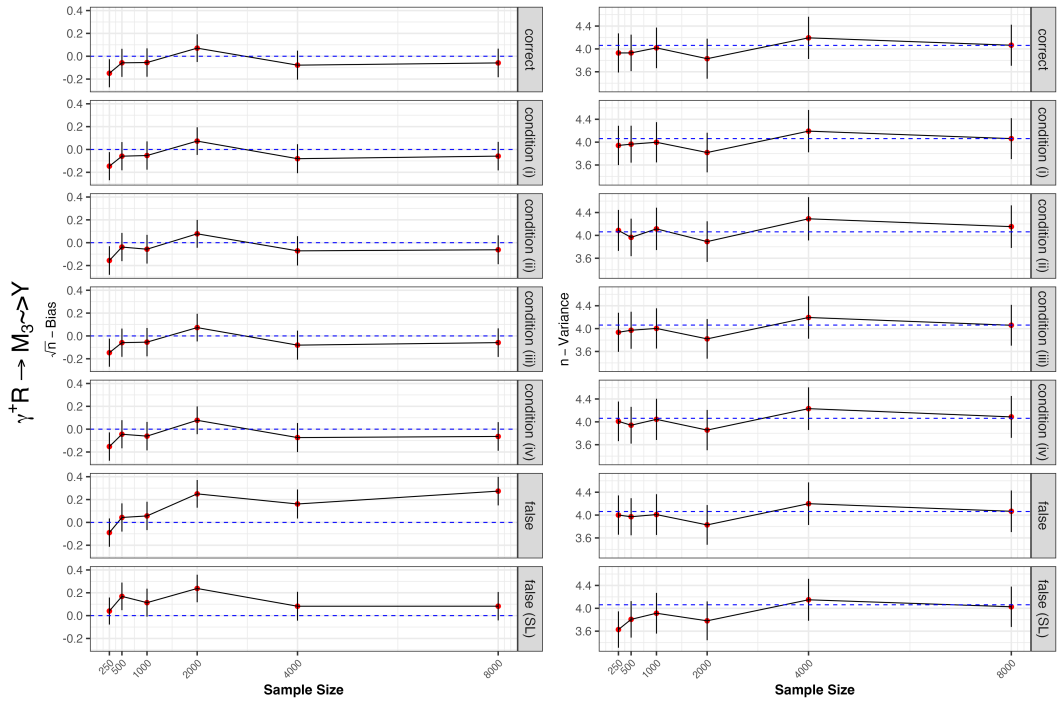
The one-step estimators under each condition are derived by combining estimated nuisance functions from both Q and Q^{false}. In addition, we also consider two additional scenarios where all nuisance functions are misspecified using both GLMs and super learner.

Figure 5.1 illustrates that the one-step estimators achieve root-n consistency under correct model specification and specific model misspecification conditions, underscoring their robustness. In contrast, estimators based solely on misspecified GLM nuisance estimates do not maintain the root-n-scaled bias property. Notably, the super learner offers a significant advantage, achieving root-n consistency even with misspecified models in large samples.


γ+R → Y
√n – Bias
n – Variance
correct
condition (i)
condition (ii)
condition (iii)
false
false (SL)
Sample Size


γ+R → M1~>Y
√n – Bias
n – Variance
correct
condition (i)
condition (ii)
condition (iii)
false
false (SL)
Sample Size


$\gamma^+R \rightarrow M_2 \rightsquigarrow Y$
$\sqrt{n}$ – Bias
n – Variance
correct
condition (i)
condition (ii)
condition (iii)
condition (iv)
false
false (SL)
Sample Size


$\gamma^+R \rightarrow M_3 \rightsquigarrow Y$
$\sqrt{n}$ – Bias
n – Variance
correct
condition (i)
condition (ii)
condition (iii)
condition (iv)
false
false (SL)
Sample Size

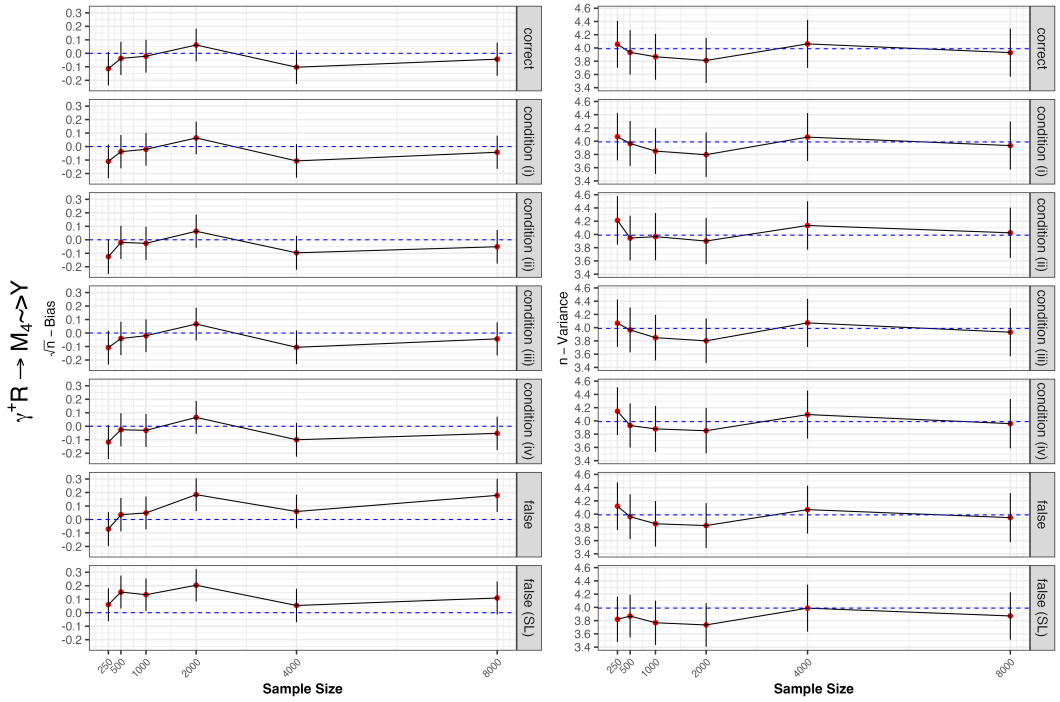


Figure 5.1: Simulation results validating the $\sqrt{n}$-consistency behaviors when the nuisance functions are misspecified under different conditions. "false" refers to estimators that utilize all misspecified GLM nuisance functions from Q^{false}, while "false (SL)" refers to estimators that rely on all nuisance functions from Q^{SL}.

5.2 Simulation 2: Finite sample performance

In this simulation, we evaluate the finite-sample performance of our estimators in the ratio of scaled geometric mean ($\rho^{+}_{R\to Y}$ and $\rho^{+}_{R\to M_k \rightsquigarrow Y}$), using both super learner and GLM. We generated data with ten covariates, one binary treatment, four ordered multivariate mediators, and a zero-inflated, right-skewed outcome, incorporating strong nonlinearities via the following models as a complex data structure

$(X_1, \ldots, X_{10}, R, M_{11}, M_{12}, M_{21}, M_{22}, M_{31}, M_{32}, M_{41}, M_{42}, Y)$:

$$X_i \overset{iid}{\sim} \text{Uniform}(0,1), i \in \{1,\ldots,8\}, \quad X_9 \sim \text{Bernoulli}(0.646), \quad X_{10} \sim \text{Bernoulli}(0.599)$$

$$Z = \begin{bmatrix} X_1^{0.5} & X_2^2 & X_3^3 & \exp(X_4) & |\log(X_5+0.5)| & \sin(X_6) & \cos(X_7-0.5) & X_8 & X_9 & X_{10} \end{bmatrix}$$

$$R \sim \text{Bernoulli}(\text{expit}(V_R \begin{bmatrix} 1 & Z \end{bmatrix}^T))$$

$$M_1 = \begin{bmatrix} M_{11} & M_{12} \end{bmatrix}, M_{12} \sim \text{Bernoulli}(\text{expit}(M_{12}^*)),$$
$$\begin{bmatrix} M_{11} \\ M_{12}^* \end{bmatrix} \sim \mathcal{N}\left(V_{M_1} \begin{bmatrix} 1 & R & Z & RZ_{9-10} \end{bmatrix}^T, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),$$

$$M_2 = \begin{bmatrix} M_{21} & M_{22} \end{bmatrix}, M_{22} \sim \text{Bernoulli}(\text{expit}(M_{22}^*)),$$
$$\begin{bmatrix} M_{21} \\ M_{22}^* \end{bmatrix} \sim \mathcal{N}\left(V_{M_2} \begin{bmatrix} 1 & R & Z & RZ_{1-4} & M_1 \end{bmatrix}^T, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),$$

$$M_3 = \begin{bmatrix} M_{31} & M_{32} \end{bmatrix}, M_{32} \sim \text{Bernoulli}(\text{expit}(M_{32}^*)),$$
$$\begin{bmatrix} M_{31} \\ M_{32}^* \end{bmatrix} \sim \mathcal{N}\left(V_{M_3} \begin{bmatrix} 1 & R & Z & M_1 & M_{21} & RM_{22} \end{bmatrix}^T, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),$$

$$M_4 = \begin{bmatrix} M_{41} & M_{42} \end{bmatrix}, M_{42} \sim \text{Bernoulli}(\text{expit}(M_{42}^*)),$$
$$\begin{bmatrix} M_{41} \\ M_{42}^* \end{bmatrix} \sim \mathcal{N}\left(V_{M_4} \begin{bmatrix} 1 & R & Z & M_1 & M_2 & M_3 \end{bmatrix}^T, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),$$

$$Y^* = V_Y \begin{bmatrix} 1 & R & Z & M_1 & M_2 & M_3 & M_{41} & M_{41}M_{42} \end{bmatrix}^T - 5,$$

$$\mathbb{I}(Y > 0) \sim \text{Bernoulli}(\text{expit}(Y^*)),$$

$$Y|Y > 0 \sim \text{LogNormal}(\log\mu = 0.15Y^*, \log sd = 0). \tag{5.4}$$

where

$Z_{1-4} = [Z_1, Z_2, Z_3, Z_4], Z_{9-10} = [Z_9, Z_{10}]$,

$V_R = [0.60, 0.39, -0.12, 0.08, -0.50, -0.46, 0.28, 0.25, 0.14, -0.45, -0.16]$,

$$V_{M_1} = \begin{bmatrix} V_{M_{11}} \\ V_{M_{12}} \end{bmatrix}, V_{M_2} = \begin{bmatrix} V_{M_{21}} \\ V_{M_{22}} \end{bmatrix}, V_{M_3} = \begin{bmatrix} V_{M_{31}} \\ V_{M_{32}} \end{bmatrix}, V_{M_4} = \begin{bmatrix} V_{M_{41}} \\ V_{M_{42}} \end{bmatrix},$$

$V_{M_{11}} = [0.76, 0.25, -0.23, 0.34, -0.21, 0.18, 0.41, -0.32, 0.15, 0.99, -0.03, 0.41, 0.26, 0.85]$,

$V_{M_{12}} = [-0.21, 0.27, -0.13, 0.22, -0.50, -0.35, -0.04, 0.09, 0.38, 0.29, -0.42, 0.82, 0.18, 0.06]$,

$V_{M_{21}} = [0.06, -0.19, -0.36, 0.10, -0.15, 0.13, -0.33, -0.50, -0.07, 0.28, -0.34, -0.09, -0.04, -0.06, 0.06, 0.04, -0.13, -0.16]$,

$V_{M_{22}} = [0.24, -0.15, -0.20, -0.06, -0.19, 0.23, -0.27, -0.50, 0.03, 0.44, -0.15, -0.07, -0.17, 0.03, 0.06, 0.09, -0.18, -0.11]$,

$V_{M_{31}} = [0.28, 0.27, 0.25, -0.92, -0.18, 0.13, -0.04, 0.96, 0.40, 0.37, 0.07, 0.25, 0.04, 0.15, -0.89, -0.65]$,

$V_{M_{32}} = [-0.99, 0.33, 0.11, -0.89, -0.11, 0.12, -0.25, 0.28, 0.20, 0.15, -0.29, 0.17, 0.06, 0.12, -0.84, -0.94]$,

$V_{M_{41}} = [-0.26, 0.73, 0.33, -0.82, 0.22, 0.11, 0.53, -0.28, -0.29, 0.25, 0.14, -0.32, 0.16, 0.19, -0.58, -0.44, 0.85, 0.74]$,

$V_{M_{42}} = [-0.87, 0.51, -0.29, -0.93, 0.31, 0.21, 0.88, -0.87, -0.46, 0.07, 0.08, -0.69, 0.14, 0.04, -0.82, -0.63, 0.78, 0.36]$,

$V_Y = [-0.76, 0.96, 0.36, 0.49, 0.64, 0.78, 0.68, -0.24, -0.64, 0.60, 0.22, 0.43, 0.69, 0.72, -0.93, -0.81, 0.94, 0.84, 0.72, 0.65]$.

In details, covariates X consist of 10 dimensions, including 8 continuous variables and 2 binary variables. Latent variables Z are transformations of X designed to introduce greater complexity and nonlinearity into the data generation process. We have four ordered mediators and each with two dimensions: one continuous and one binary variable. The binary dimension of each mediator is generated using latent variable M_{i2}^* ($i \in \{1, 2, 3, 4\}$), accounting for internal correlations within each mediator. The outcome Y is generated as a zero-inflated and right-skewed distribution, where a binomial distribution determines whether $Y = 0$, and a lognormal distribution is used to generate positive values of Y. This simulation runs for sample sizes of 1000, 2000, 4000, and 8000, with 1000 replications per scenario.

We assess our estimators using bias, standard deviation (SD), mean squared error (MSE), 95% confidence interval (CI) coverage, and average CI width. We use the same candidate learners as in the empirical analysis, with GLM-based nuisance estimation limited to simple linear or logistic regressions (without interactions or higher-order terms). Table 5.1 shows that the super learner approach yields low bias, reduced SD and MSE, and good coverage, whereas GLM-based estimators suffer from large bias and poor coverage.

These findings confirm the reliability of our empirical results and highlight the

super learner's advantage in capturing complex relationships, particularly in large-sample settings. In addition, we present a supplementary simulation with a simplified data structure in Appendix D. Table D.1 shows that GLM performs well under mild nonlinearities, making it a valuable, computationally efficient option in simpler settings.

Table 5.1: Comparative performance of one-step estimator using super learner (SL) vs. GLM in complex data structure

	Bias		SD		MSE		Coverage Rate		CI width	
sample size	SL	GLM	SL	GLM	SL	GLM	SL	GLM	SL	GLM
$\rho^+_{R\to M_1\rightsquigarrow Y}$										
1000	-0.004	-0.004	0.041	0.050	0.002	0.003	0.866	0.949	0.133	0.193
2000	0.000	-0.002	0.032	0.037	0.001	0.001	0.889	0.935	0.106	0.137
4000	0.000	-0.004	0.024	0.024	0.001	0.001	0.910	0.957	0.083	0.096
8000	0.001	-0.004	0.018	0.017	0.000	0.000	0.925	0.952	0.065	0.068
$\rho^+_{R\to M_2\rightsquigarrow Y}$										
1000	0.007	0.012	0.041	0.045	0.002	0.002	0.902	0.966	0.139	0.184
2000	0.003	0.008	0.031	0.033	0.001	0.001	0.902	0.957	0.104	0.128
4000	0.003	0.007	0.023	0.023	0.001	0.001	0.910	0.943	0.078	0.090
8000	0.001	0.007	0.016	0.015	0.000	0.000	0.937	0.947	0.058	0.064
$\rho^+_{R\to M_3\rightsquigarrow Y}$										
1000	-0.001	0.013	0.025	0.033	0.001	0.001	0.833	0.947	0.070	0.127
2000	-0.001	0.012	0.017	0.024	0.000	0.001	0.868	0.929	0.054	0.089
4000	-0.001	0.012	0.012	0.016	0.000	0.000	0.907	0.909	0.041	0.063
8000	0.000	0.013	0.009	0.011	0.000	0.000	0.916	0.816	0.031	0.045
$\rho^+_{R\to M_4\to Y}$										
1000	-0.003	0.025	0.018	0.031	0.000	0.002	0.823	0.915	0.051	0.124
2000	-0.002	0.025	0.013	0.023	0.000	0.001	0.866	0.839	0.040	0.088
4000	0.000	0.025	0.010	0.017	0.000	0.001	0.868	0.665	0.031	0.063
8000	0.000	0.026	0.007	0.012	0.000	0.001	0.904	0.371	0.025	0.044
$\rho^+_{R\to Y}$										
1000	-0.004	-0.001	0.006	0.010	0.000	0.000	0.817	0.948	0.020	0.036
2000	-0.002	-0.001	0.005	0.007	0.000	0.000	0.893	0.947	0.016	0.027
4000	-0.001	-0.001	0.003	0.005	0.000	0.000	0.925	0.962	0.012	0.019
8000	0.000	-0.001	0.003	0.004	0.000	0.000	0.922	0.933	0.010	0.013

6 Discussion

This study examines racial disparities in U.S. healthcare expenditures using a causal path-specific effects framework. We evaluate how socioeconomic status (SES), insurance access, health behaviors, and health status contribute to these disparities. Our approach integrates flexible estimation with data-adaptive modeling and is implemented via the `flexPaths` R package, providing a methodological template for pathway analysis in disparities research.

Our path-specific decomposition offers valuable insights for policy. The strong SES-mediated disparities indicate that investments in education and income support may help reduce healthcare inequities. Similarly, the pronounced role of insurance access disparities suggests that targeted expansions of coverage for groups with historically high uninsurance rates could substantially improve expenditure equity. By pinpointing the key pathways driving disparities, our analysis provides a data-driven basis for targeted policy interventions addressing structural inequities.

Cost data are widely used to predict healthcare needs, yet they often mirror underlying disparities. Prior work has shown that algorithms relying solely on cost data may underestimate Black patients' needs relative to White patients, potentially deprioritizing those most in need [62]. This highlights the necessity for fairness-aware adjustments, with counterfactual and causal reasoning emerging as essential tools for quantifying fairness [58, 59, 96, 49, 23, 17]. Our path-specific decomposition reveals that the effect of race on health spending is significantly mediated by factors such as SES and insurance access. If predictive models overlook these mediated disparities, they risk reinforcing existing inequities. A fairness-aware algorithm could constrain either the direct or indirect effects of race, thereby aligning predictions more closely with equitable healthcare access.

Despite its strengths, this study has several limitations. First, the reliance on self-reported data introduces potential reporting biases, as participants may misreport conditions due to recall errors or social desirability. Although cross-referencing with clinical records could mitigate this issue, such data are often difficult to access. Second, selection bias is a concern if marginalized populations are underrepresented; future research should assess this bias and incorporate more diverse data sources to capture healthcare access comprehensively. Third, the causal interpretation of race is inherently challenging because race is a social construct and not directly manipulable. Although our framework focuses on actionable mediators such as SES and insurance access, the underlying assumptions of causal mediation—especially the cross-world counterfactual independence—are difficult to verify empirically. These limitations warrant cautious interpretation and call for further methodological refinement.

Future studies should extend our findings by broadening the scope of mediators to isolate specific pathways—for example, examining the direct link between SES and expenditures independently of its downstream effects. Additionally, sensitivity analyses to assess the robustness of our causal assumptions, particularly concerning unmeasured confounding, are essential. Further research could also explore modifications to the mediation framework to capture dynamic changes over time and incorporate alternative measures of health outcomes. These efforts will enhance our understanding of the multifaceted mechanisms driving healthcare inequities.

A Proofs

A.1 Identification claims

Under the ignorability assumptions from Chapter 3, the estimands in Theorem 3.2.1 are identified via an identification of the counterfactual mean $\mathbb{E}(Y(r_0, r_1, r_2, r_3, r_4))$, as follows:

$$
\begin{aligned}
&\mathbb{E}(Y(r_0, r_1, r_2, r_3, r_4)) \\
&= \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, M_4(r_4, \overline{m}_3) = m_4, X\Big] \\
&\qquad dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, X\big) \\
&\qquad dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, X\big) \\
&\qquad dP\big(M_2(r_2, m_1) = m_2 \mid M_1(r_1) = m_1, X\big) dP\big(M_1(r_1) = m_1 \mid X\big) dP(x) \\
&\overset{A3.1}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, M_4(r_4, \overline{m}_3) = m_4, R = r_0, X\Big] \\
&\qquad dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, R = r_4, X\big) \\
&\qquad dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, R = r_3, X\big) \\
&\qquad dP\big(M_2(r_2, m_1) = m_2 \mid M_1(r_1) = m_1, R = r_2, X\big) dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x) \\
&\overset{A3.5}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, R = r_0, X\Big] \\
&\qquad dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, M_3(r_3, \overline{m}_2) = m_3, R = r_4, X\big) \\
&\qquad dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, R = r_3, X\big) \\
&\qquad dP\big(M_2(r_2, m_1) = m_2 \mid M_1(r_1) = m_1, R = r_2, X\big) dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x) \\
&\overset{A3.4}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, R = r_0, X\Big] \\
&\qquad dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, R = r_4, X\big) \\
&\qquad dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1(r_1) = m_1, M_2(r_2, m_1) = m_2, R = r_3, X\big) \\
&\qquad dP\big(M_2(r_2, m_1) = m_2 \mid M_1(r_1) = m_1, R = r_2, X\big) dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x) \\
&\overset{A3.3}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1(r_1) = m_1, R = r_0, X\Big] dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1(r_1) = m_1, R = r_4, X\big) \\
&\qquad dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1(r_1) = m_1, R = r_3, X\big) dP\big(M_2(r_2, m_1) = m_2 \mid M_1(r_1) = m_1, R = r_2, X\big) \\
&\qquad dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)
\end{aligned}
$$

$$\overset{A3.2}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid R = r_0, X\Big] dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid R = r_4, X\big) dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid R = r_3, X\big)$$
$$dP\big(M_2(r_2, m_1) = m_2 \mid R = r_2, X\big) dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)$$
$$\overset{A3.2\&A1}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1 = m_1, R = r_0, X\Big] dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1 = m_1, R = r_4, X\big)$$
$$dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1 = m_1, R = r_3, X\big) dP\big(M_2(r_2, m_1) = m_2 \mid M_1 = m_1, R = r_2, X\big)$$
$$dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)$$
$$\overset{A3.3\&A1}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1 = m_1, M_2 = m_2, R = r_0, X\Big] dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1 = m_1, M_2 = m_2, R = r_4, X\big)$$
$$dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1 = m_1, M_2 = m_2, R = r_3, X\big) dP\big(M_2(r_2, m_1) = m_2 \mid M_1 = m_1, R = r_2, X\big)$$
$$dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)$$
$$\overset{A3.4\&A1}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1 = m_1, M_2 = m_2, M_3 = m_3, R = r_0, X\Big]$$
$$dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1 = m_1, M_2 = m_2, M_3 = m_3, R = r_4, X\big)$$
$$dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1 = m_1, M_2 = m_2, R = r_3, X\big)$$
$$dP\big(M_2(r_2, m_1) = m_2 \mid M_1 = m_1, R = r_2, X\big)$$
$$dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)$$
$$\overset{A3.5\&A1}{=} \int \mathbb{E}\Big[Y(r_0, \overline{m}_4) \mid M_1 = m_1, M_2 = m_2, M_3 = m_3, M_4 = m_4, R = r_0, X\Big]$$
$$dP\big(M_4(r_4, \overline{m}_3) = m_4 \mid M_1 = m_1, M_2 = m_2, M_3 = m_3, R = r_4, X\big)$$
$$dP\big(M_3(r_3, \overline{m}_2) = m_3 \mid M_1 = m_1, M_2 = m_2, R = r_3, X\big)$$
$$dP\big(M_2(r_2, m_1) = m_2 \mid M_1 = m_1, R = r_2, X\big)$$
$$dP\big(M_1(r_1) = m_1 \mid R = r_1, X\big) dP(x)$$
$$\overset{A1}{=} \int y dP(y \mid r_0, \overline{m}_4, x) dP(m_4 \mid r_4, \overline{m}_3, x) dP(m_3 \mid r_3, \overline{m}_2, x) dP(m_2 \mid r_2, m_1, x) dP(m_1 \mid r_1, x) dP(x) \ .$$

These derivations yield the identification functionals for the estimands in Theorem 3.2.1.

A.2 Estimation claims

Let $o = (x, r, \overline{m}_4, y)$ denote the vector values of $O = (X, R, \overline{M}_4, Y)$.

First, note that by the Bayes' rule, we can write:

$$\begin{aligned}\frac{p(m_k \mid \overline{m}_{k-1}, R=1, x)}{p(m_k \mid \overline{m}_{k-1}, R=0, x)} &= \frac{p(R=1 \mid \overline{m}_k, x)p(m_k \mid \overline{m}_{k-1}, x)/p(R=1 \mid \overline{m}_{k-1}, x)}{p(R=0 \mid \overline{m}_k, x)p(m_k \mid \overline{m}_{k-1}, x)/p(R=0 \mid \overline{m}_{k-1}, x)} \\ &= \frac{g_k(\overline{m}_k, x)}{1-g_k(\overline{m}_k, x)} \frac{1-g_{k-1}(\overline{m}_{k-1}, x)}{g_{k-1}(\overline{m}_{k-1}, x)} \ .\end{aligned} \tag{A.1}$$

- EIF derivation for $\gamma_{R\to Y}$:

$$\begin{aligned}&\frac{\partial}{\partial \varepsilon}\gamma_{R\to Y}(P_\varepsilon)\Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial \varepsilon}\int y dP_\varepsilon(y \mid \overline{m}_4, R=1, x) dP_\varepsilon(\overline{m}_4 \mid R=0, x) dP_\varepsilon(x)\Big|_{\varepsilon=0} \\ &= \int y S(y \mid \overline{m}_4, R=1, x) dP(y \mid \overline{m}_4, R=1, x) dP(\overline{m}_4 \mid R=0, x) dP(x) \qquad (1) \\ &+ \int y S(\overline{m}_4 \mid R=0, x) dP(y \mid \overline{m}_4, R=1, x) dP(\overline{m}_4 \mid R=0, x) dP(x) \qquad (2) \\ &+ \int y S(x) dP(y \mid \overline{m}_4, R=1, x) dP(\overline{m}_4 \mid R=0, x) dP(x) \ . \qquad (3)\end{aligned}$$

Line (1) simplifies to:

$$\begin{aligned}&\int y S(y \mid \overline{m}_4, R=1, x) dP(y \mid \overline{m}_4, R=1, x) dP(\overline{m}_4 \mid R=0, x) dP(x) \\ &= \int \frac{\mathbb{I}(R=1)}{p(R=1 \mid x)} \frac{p(\overline{m}_4 \mid R=0, x)}{p(\overline{m}_4 \mid R=1, x)} y S(y \mid \overline{m}_4, R, x) dP(y, \overline{m}_4, R, x) \\ &\overset{A.1}{=} \int \frac{\mathbb{I}(R=1)}{1-\pi(x)} \frac{1-g_4(\overline{m}_4, x)}{g_4(\overline{m}_4, x)} \big(y - \mu_4(\overline{m}_4, R=1, x)\big) S(o) dP(o) \ .\end{aligned}$$

Line (2) simplifies to:

$$\begin{aligned}&\int y S(\overline{m}_4 \mid R=0, x) dP(y \mid \overline{m}_4, R=1, x) dP(\overline{m}_4 \mid R=0, x) dP(x) \\ &= \int \frac{\mathbb{I}(R=0)}{p(R=0 \mid x)} \mu_4(\overline{m}_4, R=1, x) S(\overline{m}_4 \mid R, x) dP(\overline{m}_4, R, x) \\ &= \int \frac{\mathbb{I}(R=0)}{1-\pi(x)} \big(\mu_4(\overline{m}_4, R=1, x) - \mathcal{C}_{\mu_4}(R=0, x)\big) S(o) dP(o) \ .\end{aligned}$$

Line (3) simplifies to:

$$\begin{aligned}
&\int yS(x)dP(y \mid \overline{m}_4, R=1, x)dP(\overline{m}_4 \mid R=0, x)dP(x) \\
&= \int \mathcal{C}_{\mu_4}(R=0, x)S(x)dP(o) \\
&= \int \big(\mathcal{C}_{\mu_4}(R=0, x) - \gamma_{R\to Y}\big)S(o)dP(o) \ .
\end{aligned}$$

Therefore, the EIF for $\gamma_{R\to Y}$, denoted by $\Phi_{\gamma_{R\to Y}}(Q)$, is given as follows:

$$\begin{aligned}
\Phi_{\gamma_{R\to Y}}(Q)(O) =& \frac{R}{1-\pi(X)}\frac{1-g_4(\overline{M}_4, X)}{g_4(\overline{M}_4, X)}\left\{Y - \mu_4(\overline{M}_4, R=1, X)\right\} \\
&+ \frac{1-R}{1-\pi(X)}\left\{\mu_4(\overline{M}_4, R=1, X) - \mathcal{C}_{\mu_4}(R=0, X)\right\} + \mathcal{C}_{\mu_4}(R=0, X) - \gamma_{R\to Y} \ .
\end{aligned} \tag{A.2}$$

• EIF derivation for $\gamma_{R\to M_k\rightsquigarrow Y}$, $k = 2, 3, 4$, where:

$$\gamma_{R\to M_k\rightsquigarrow Y} = \int ydP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \ .$$

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\gamma_{R\to M_k\rightsquigarrow Y}(P_\varepsilon)\Big|_{\varepsilon=0} \\
&= \frac{\partial}{\partial\varepsilon}\int ydP_\varepsilon(y \mid \overline{m}_k, R=0, x)dP_\varepsilon(m_k \mid \overline{m}_{k-1}, R=1, x)dP_\varepsilon(\overline{m}_{k-1} \mid R=0, x)dP_\varepsilon(x)\Big|_{\varepsilon=0} \\
&= \int yS(y \mid \overline{m}_k, R=0, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \qquad (1) \\
&+ \int yS(m_k \mid \overline{m}_{k-1}, R=1, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \qquad (2) \\
&+ \int yS(\overline{m}_{k-1} \mid R=0, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \qquad (3) \\
&+ \int yS(x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \ . \qquad (4)
\end{aligned}$$

Line (1) simplifies to:

$$\begin{aligned}
&\int yS(y \mid \overline{m}_k, R=0, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \\
&= \int \frac{\mathbb{I}(R=0)}{1-\pi(x)} \frac{p(m_k \mid \overline{m}_{k-1}, R=1, x)}{p(m_k \mid \overline{m}_{k-1}, R=0, x)} yS(y \mid \overline{m}_k, R, x)dP(y, \overline{m}_k, R, x) \\
&= \int \frac{\mathbb{I}(R=0)}{1-\pi(x)} \frac{p(m_k \mid \overline{m}_{k-1}, R=1, x)}{p(m_k \mid \overline{m}_{k-1}, R=0, x)} \big(y - \mu_k(\overline{m}_k, R=0, x)\big) S(o)dP(o) \\
&\overset{A.1}{=} \int \frac{\mathbb{I}(R=0)}{1-\pi(x)} \frac{g_k(\overline{m}_k, x)}{1-g_k(\overline{m}_k, x)} \frac{1-g_{k-1}(\overline{m}_{k-1}, x)}{g_{k-1}(\overline{m}_{k-1}, x)} \big(y - \mu_k(\overline{m}_k, R=0, x)\big) S(o)dP(o) \ .
\end{aligned}$$

Line (2) simplifies to:

$$\begin{aligned}
&\int yS(m_k \mid \overline{m}_{k-1}, R=1, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \\
&= \int \frac{\mathbb{I}(R=1)}{p(R=1 \mid x)} \frac{p(\overline{m}_{k-1} \mid R=0, x)}{p(\overline{m}_{k-1} \mid R=1, x)} \mu_k(\overline{m}_k, R=0, x) S(m_k \mid \overline{m}_{k-1}, R, x)dP(\overline{m}_k, R, x) \\
&\overset{A.1}{=} \int \frac{\mathbb{I}(R=1)}{1-\pi(x)} \frac{1-g_{k-1}(\overline{m}_{k-1}, x)}{g_{k-1}(\overline{m}_{k-1}, x)} \big(\mu_k(\overline{m}_k, R=0, x) - \mathcal{B}_k(\overline{m}_{k-1}, R=1, x)\big) S(o)dP(o) \ .
\end{aligned}$$

Line (3) simplifies to:

$$\begin{aligned}
&\int yS(\overline{m}_{k-1} \mid R=0, x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \\
&= \int \frac{\mathbb{I}(R=0)}{p(R=0 \mid x)} \mathcal{B}_k(\overline{m}_{k-1}, R=1, x) S(\overline{m}_{k-1} \mid R, x)dP(\overline{m}_{k-1}, R, x) \\
&= \int \frac{\mathbb{I}(R=0)}{1-\pi(x)} \big(\mathcal{B}_k(\overline{m}_{k-1}, R=1, x) - \mathcal{C}_{\mathcal{B}_k}(R=0, x)\big) S(o)dP(o) \ .
\end{aligned}$$

Line (4) simplifies to:

$$\begin{aligned}
&\int yS(x)dP(y \mid \overline{m}_k, R=0, x)dP(m_k \mid \overline{m}_{k-1}, R=1, x)dP(\overline{m}_{k-1} \mid R=0, x)dP(x) \\
&= \int \mathcal{C}_{\mathcal{B}_k}(R=0, x) S(x)dP(x) \\
&= \int \big(\mathcal{C}_{\mathcal{B}_k}(R=0, x) - \gamma_{R \to M_k \rightsquigarrow Y}\big) S(o)dP(o) \ .
\end{aligned}$$

Therefore, the EIF for $\gamma_{R\to M_k\rightsquigarrow Y}$, denoted by $\Phi_{\gamma_{R\to M_k\rightsquigarrow Y}}(Q)$, is given by:

$$\begin{aligned}\Phi_{\gamma_{R\to M_k\rightsquigarrow Y}}(Q)(O) = &\frac{1-R}{1-\pi(X)}\frac{g_k(\overline{M}_k,X)}{1-g_k(\overline{M}_k,X)}\frac{1-g_{k-1}(\overline{M}_{k-1},X)}{g_{k-1}(\overline{M}_{k-1},X)}\left\{Y-\mu_k(\overline{M}_k,R=0,X)\right\}\\ &+\frac{R}{1-\pi(X)}\frac{1-g_{k-1}(\overline{M}_{k-1},X)}{g_{k-1}(\overline{M}_{k-1},X)}\left\{\mu_k(\overline{M}_k,R=0,X)-\mathcal{B}_k(\overline{M}_{k-1},R=1,X)\right\}\\ &+\frac{1-R}{1-\pi(x)}\left\{\mathcal{B}_k(\overline{m}_{k-1},R=1,x)-\mathcal{C}_{\mathcal{B}_k}(R=0,x)\right\}\\ &+\mathcal{C}_{\mathcal{B}_k}(R=0,x)-\gamma_{R\to M_k\rightsquigarrow Y}\ .\end{aligned} \tag{A.3}$$

• EIF derivation for $\gamma_{R\to M_1\rightsquigarrow Y}$, where

$$\gamma_{R\to M_1\rightsquigarrow Y}=\int y dP(y\mid m_1,R=0,x)dP(m_1\mid R=1,x)dP(x)\ .$$

$$\begin{aligned}&\frac{\partial}{\partial\varepsilon}\gamma_{R\to M_1\rightsquigarrow Y}(P_\varepsilon)\Big|_{\varepsilon=0}\\ &=\frac{\partial}{\partial\varepsilon}\int y dP_\varepsilon(y\mid m_1,R=0,x)dP_\varepsilon(m_1\mid R=1,x)dP_\varepsilon(x)\Big|_{\varepsilon=0}\\ &=\int yS(y\mid m_1,R=0,x)dP(y\mid m_1,R=0,x)dP(m_1\mid R=1,x)dP(x) &(1)\\ &+\int yS(m_1\mid R=1,x)dP(y\mid m_1,R=0,x)dP(m_1\mid R=1,x)dP(x) &(2)\\ &+\int yS(x)dP(y\mid m_1,R=0,x)dP(m_1\mid R=1,x)dP(x)\ . &(3)\end{aligned}$$

Line (1) simplifies to:

$$\begin{aligned}
&\int yS(y \mid m_1, R=0, x)dP(y \mid m_1, R=0, x)dP(m_1 \mid R=1, x)dP(x) \\
&= \int \frac{\mathbb{I}(R=0)}{p(R=0 \mid x)} \frac{p(m_1 \mid R=1, x)}{p(m_1 \mid R=0, x)} yS(y \mid m_1, R=0, x)dP(y, m_1, R=0, x) \\
&= \int \frac{\mathbb{I}(R=0)}{\pi(x)} \frac{p(m_1 \mid R=1, x)}{p(m_1 \mid R=0, x)} yS(y \mid m_1, R, x)dP(y, m_1, R, x) \\
&\overset{A.1}{=} \int \frac{\mathbb{I}(R=0)}{\pi(x)} \frac{g_1(m_1, x)}{1-g_1(m_1, x)} \big(y - \mu_1(m_1, R=0, x)\big) S(o)dP(o) \ .
\end{aligned}$$

Line (2) simplifies to:

$$\begin{aligned}
&\int yS(m_1 \mid R=1, x)dP(y \mid m_1, R=0, x)dP(m_1 \mid R=1, x)dP(x) \\
&= \int \frac{\mathbb{I}(R=1)}{p(R=1 \mid x)} \mu_1(m_1, R=0, x) S(m_1 \mid R, x)dP(m_1, R, x) \\
&= \int \frac{\mathbb{I}(R=1)}{\pi(x)} \big(\mu_1(m_1, R=0, x) - \mathcal{C}_{\mu_1}(R=1, x)\big) S(o)dP(o) \ .
\end{aligned}$$

Line (3) simplifies to:

$$\begin{aligned}
&\int yS(x)dP(y \mid m_1, R=0, x)dP(m_1 \mid R=1, x)dP(x) \\
&= \int \mathcal{C}_{\mu_1}(R=1, x) S(x)dP(x) \\
&= \int \big(\mathcal{C}_{\mu_1}(R=1, x) - \gamma_{R \to M_1 \rightsquigarrow Y}\big) S(x)dP(x) \ .
\end{aligned}$$

Therefore, the EIF for $\gamma_{R \to M_1 \rightsquigarrow Y}$, denoted by $\Phi_{\gamma_{R \to M_1 \rightsquigarrow Y}}(Q)$, is given by:

$$\begin{aligned}
\Phi_{\gamma_{R \to M_1 \rightsquigarrow Y}}(Q)(O) = &\frac{1-R}{\pi(X)} \frac{g_1(M_1, X)}{1-g_1(M_1, X)} \left\{Y - \mu_1(M_1, R=0, X)\right\}) \\
&+ \frac{R}{\pi(X)} \left\{\mu_1(M_1, R=0, X) - \mathcal{C}_{\mu_1}(R=1, X)\right\} + \mathcal{C}_{\mu_1}(R=1, x) - \gamma_{R \to M_1 \rightsquigarrow Y} \ .
\end{aligned} \tag{A.4}$$

Due to the identities $g_0(M_0, X) = \pi(X)$ and $\mathcal{B}_1(R = 1, x) = \mathcal{C}_{\mathcal{B}_1}(R = 1, x)$, the EIF for $\gamma_{R \to M_1 \rightsquigarrow Y}$ can be incorporated into the expression for $\gamma_{R \to M_k \rightsquigarrow Y}$ for $k = 2, 3, 4$.

A.3 Inference claims

In Theorem 3.3.1 and Corollary 3.3.2, certain regularity conditions are required for the empirical process term to be negligible, i.e., $(P_n - P)(\Phi(\hat{Q}) - \Phi(Q)) = o_P(n^{-1/2}))$. These conditions are as follows:

1. $\Phi(\hat{Q}) - \Phi(Q)$ belongs to a P-Donsker class with probability tending to 1, and
2. $\Phi(\hat{Q})$ is $L^2(P)$-consistent: $P\{\Phi(\hat{Q}) - \Phi(Q)\}^2 = o_P(1)$.

The first condition can be relaxed using sample-splitting procedures [22]. Additionally, we require, for $\delta > 0$: $\delta < \hat{\pi} < 1 - \delta$ and $\delta < \hat{g}_k < 1 - \delta$, $k = 1, 2, 3, 4$.

It remains to derive the remainder terms for $\gamma^+_{R \to Y}(\hat{Q})$ and $\gamma^+_{R \to M_k \rightsquigarrow Y}(\hat{Q})$, denoted by $R_{2,\gamma_{R \to Y}}(\hat{Q}, Q)$ and $R_{2,\gamma_{R \to M_k \rightsquigarrow Y}}(\hat{Q}, Q)$, respectively. In below, we show these remainder terms are: ($\pi \equiv g_0$ and $\mathcal{B}_1 \equiv \mathcal{C}_{\mathcal{B}_1}$)

$$R_{2,\gamma_{R \to Y}}(\hat{Q}, Q) = P\left[\frac{1}{1-\hat{\pi}}\frac{1}{\hat{g}_4}(\hat{g}_4 - g_4)(\hat{\mu}_4 - \mu_4) + \frac{1}{1-\hat{\pi}}(\pi - \hat{\pi})(\hat{\mathcal{C}}_{\mu_4} - \mathcal{C}_{\mu_4})\right], \tag{A.5}$$

$$\begin{aligned} R_{2,\gamma_{R \to M_k \rightsquigarrow Y}}(\hat{Q}, Q) = P\Big[&\frac{1}{1-\hat{\pi}}\frac{1}{\hat{g}_{k-1}}\Big\{\frac{1-\hat{g}_{k-1}}{1-\hat{g}_k}(g_k - \hat{g}_k)(\hat{\mu}_k - \mu_k) + (\hat{g}_{k-1} - g_{k-1})(\hat{\mathcal{B}}_k - \mathcal{B}_k)\Big\} \\ &+ \frac{1}{1-\hat{\pi}}(\pi - \hat{\pi})(\hat{\mathcal{C}}_{\mathcal{B}_k} - \mathcal{C}_{\mathcal{B}_k})\Big]\ ,\ k = 1, 2, 3, 4\ . \end{aligned} \tag{A.6}$$

Note that conditions for $R_2(\hat{Q}, \hat{Q}) = o_P(n^{-1/2})$ are equivalent to each nuisance product term having an $L^2(P)$ convergence rate equal or faster than $o_P(n^{-1/2})$, with finite scaling factors.

Let $h(Q)(O) = \Phi(Q)(O) + \gamma(Q)$, and thus $\gamma^+(\hat{Q}) = P_n[h(\hat{Q})] = \frac{1}{n}\sum_{i=1}^n h(\hat{Q})(O_i)$. We propose a special set of estimated nuisance parameters $\widetilde{Q} = (\hat{\pi}, \hat{g}, \mathcal{C}, \mathcal{B}, \mu)$ where

all the outcome and sequential regression nuisances are correctly estimated. Our first step is to prove that $P[h(\widetilde{Q})] = \gamma$, where $P[h(Q)] = \int h(Q)(o)dP(o)$.

• For $\gamma_{R\to Y}$:

$$\begin{aligned} P\left[h_{\gamma_{R\to Y}}(\widetilde{Q})\right] =& P\left[\frac{R}{1-\hat{\pi}}\frac{1-\hat{g}_4}{\hat{g}_4}\underbrace{E\left(Y-\mu_4 \mid \overline{M}_4, R=1, X\right)}_{=0}\right] \\ &+ P\left[\frac{1-R}{1-\hat{\pi}}\underbrace{E\left(\mu_4 - \mathcal{C}_{\mu_4} \mid R=0, X\right)}_{=0}\right] + P\left[\mathcal{C}_{\mu_4}\right] \\ =& P\left[\mathcal{C}_{\mu_4}\right] = \gamma_{R\to Y} \ . \end{aligned} \tag{A.7}$$

• For $\gamma_{R\to M_k \rightsquigarrow y}$:

$$\begin{aligned} P[h_{\gamma_{R\to M_k\rightsquigarrow y}}(\widetilde{Q})] =& P\left[\frac{1-R}{1-\hat{\pi}}\frac{\hat{g}_k}{1-\hat{g}_k}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\underbrace{E\left(Y-\mu_k \mid \overline{M}_k, R=0, X\right)}_{=0}\right] \\ &+ P\left[\frac{R}{1-\hat{\pi}}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\underbrace{E\left(\mu_k - \mathcal{B}_k \mid \overline{M}_{k-1}, R=1, X\right)}_{=0}\right] \\ &+ P\left[\frac{1-R}{1-\hat{\pi}}\underbrace{E\left(\mathcal{B}_k - \mathcal{C}_{\mathcal{B}_k} \mid R=0, X\right)}_{=0}\right] + P\left[\mathcal{C}_{\mathcal{B}_k}\right] \\ =& P\left[\mathcal{C}_{\mathcal{B}_k}\right] = \gamma_{R\to M_k\rightsquigarrow y} \ . \end{aligned} \tag{A.8}$$

With $P[h(\widetilde{Q})] = \gamma(Q)$, the second-order remainder term can be re-written as $R_2(\hat{Q}, Q) = P[h(\hat{Q})] - \gamma(Q) = P[h(\hat{Q}) - h(\widetilde{Q})]$. Using this fact, the second-order remainder terms can be derived as follows:

$$\begin{aligned} R_{2,R\to Y}(\hat{Q}, Q) =& P\left\{\frac{R}{1-\hat{\pi}}\frac{1-\hat{g}_4}{\hat{g}_4}\left[(Y-\hat{\mu}_4)-(Y-\mu_4)\right]\right\} \\ &+ P\left\{\frac{1-R}{1-\hat{\pi}}\left[\left(\hat{\mu}_4 - \hat{\mathcal{C}}_{\mu_4}\right) - \left(\mu_4 - \mathcal{C}_{\mu_4}\right)\right]\right\} + P\left(\hat{\mathcal{C}}_{\mu_4} - \mathcal{C}_{\mu_4}\right) \end{aligned}$$

$$
\begin{aligned}
&= - P\left[\frac{g_4}{1-\pi}\frac{1-\hat{g}_4}{\hat{g}_4}\left(\hat{\mu}_4-\mu_4\right)\right] + P\left[\frac{1-g_4}{1-\hat{\pi}}\left(\hat{\mu}_4-\mu_4\right)\right] \\
&\quad - P\left[\frac{1-\pi}{1-\hat{\pi}}\left(\hat{\mathcal{C}}_{\mu_4}-\mathcal{C}_{\mu_4}\right)\right] + P\left[\hat{\mathcal{C}}_{\mu_4}-\mathcal{C}_{\mu_4}\right] \\
&= P\left[\frac{1}{1-\hat{\pi}}\frac{1}{\hat{g}_4}(\hat{g}_4-g_4)(\hat{\mu}_4-\mu_4)\right] + P\left[\frac{1}{1-\hat{\pi}}(\pi-\hat{\pi})(\hat{\mathcal{C}}_{\mu_4}-\mathcal{C}_{\mu_4})\right] . \qquad \text{(A.9)}
\end{aligned}
$$

$$
\begin{aligned}
R_{2,R\to M_k\rightsquigarrow Y}(\hat{Q},Q) =& P\left\{\frac{1-R}{1-\hat{\pi}}\frac{\hat{g}_k}{1-\hat{g}_k}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\left[(Y-\hat{\mu}_k)-(Y-\mu_k)\right]\right\} \\
&+ P\left\{\frac{R}{1-\hat{\pi}}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\left[\left(\hat{\mu}_k-\hat{\mathcal{B}}_k\right)-(\mu_k-\mathcal{B}_k)\right]\right\} \\
&+ P\left\{\frac{1-R}{1-\hat{\pi}}\left[\left(\hat{\mathcal{B}}_k-\hat{\mathcal{C}}_{\mathcal{B}_k}\right)-(\mathcal{B}_k-\mathcal{C}_{\mathcal{B}_k})\right]\right\} \\
&+ P\left\{\hat{\mathcal{C}}_{\mathcal{B}_k}-\mathcal{C}_{\mathcal{B}_k}\right\} \\
=& - P\left[\frac{1-\hat{g}_k}{1-\hat{\pi}}\frac{\hat{g}_k}{1-\hat{g}_k}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\left(\hat{\mu}_k-\mu_k\right)\right] + P\left[\frac{g_k}{1-\hat{\pi}}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\left(\hat{\mu}_k-\mu_k\right)\right] \\
&- P\left[\frac{g_{k-1}}{1-\hat{\pi}}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}\left(\hat{\mathcal{B}}_k-\mathcal{B}_k\right)\right] + P\left[\frac{1-g_{k-1}}{1-\hat{\pi}}\left(\hat{\mathcal{B}}_k-\mathcal{B}_k\right)\right] \\
&- P\left[\frac{1-\pi}{1-\hat{\pi}}\left(\hat{\mathcal{C}}_{B_k}-\mathcal{C}_{\mathcal{B}_k}\right)\right] + P\left[\hat{\mathcal{C}}_{\mathcal{B}_k}-\mathcal{C}_{\mathcal{B}_k}\right] \\
=& P\left[\frac{1}{1-\hat{\pi}}\frac{1}{1-\hat{g}_k}\frac{1-\hat{g}_{k-1}}{\hat{g}_{k-1}}(g_k-\hat{g}_k)(\hat{\mu}_k-\mu_k)\right] \\
&+ P\left[\frac{1}{1-\hat{\pi}}\frac{1}{\hat{g}_{k-1}}(\hat{g}_{k-1}-g_{k-1})(\hat{\mathcal{B}}_k-\mathcal{B}_k)\right] \\
&+ P\left[\frac{1}{1-\hat{\pi}}(\pi-\hat{\pi})(\hat{\mathcal{C}}_{\mathcal{B}_k}-\mathcal{C}_{\mathcal{B}_k})\right] , \qquad \text{(A.10)}
\end{aligned}
$$

for $k = 1, 2, 3, 4$. Note that when $k = 1$, the R_2 term reduces to:

$$
R_{2,R\to M_1\rightsquigarrow Y}(\hat{Q},Q) = P\left[\frac{1}{\hat{\pi}}\frac{1}{1-\hat{g}_1}(g_1-\hat{g}_1)(\hat{\mu}_1-\mu_1)\right] + P\left[\frac{1}{\hat{\pi}}(\hat{\pi}-\pi)(\hat{\mathcal{B}}_1-\mathcal{B}_1)\right]. \qquad \text{(A.11)}
$$

With the second-order remainder terms expressed as a sum of cross-product terms, regularity conditions are required to ensure that these terms are negligible, i.e., $o_P(n^{-1/2})$. Specifically, all denominators must be bounded away from zero. Thus, the propensity

score estimates $\hat{\pi}$ and $\hat{g}_k$ for $k = 1, 2, 3, 4$ must satisfy $0 < \hat{\pi} < 1$ and $0 < \hat{g}_k < 1$. Under this regularity assumption, the second-order remainder terms can be expressed as:

$$R_{2,R\to Y}(\hat{Q}, Q) = P\left[m_1(\hat{\pi}, \hat{g}_4) \cdot (\hat{g}_4 - g_4) \cdot (\hat{\mu}_4 - \mu_4)\right] + P\left[m_2(\hat{\pi}) \cdot (\pi - \hat{\pi}) \cdot (\hat{\mathcal{C}}_{\mu_4} - \mathcal{C}_{\mu_4})\right] , \tag{A.12}$$

$$\begin{aligned} R_{2,R\to M_k \rightsquigarrow Y}(\hat{Q}, Q) =& P\left[m_3(\hat{\pi}, \hat{g}_k, \hat{g}_{k-1}) \cdot (g_k - \hat{g}_k) \cdot (\hat{\mu}_k - \mu_k)\right] \\ &+ P\left[m_1(\hat{\pi}, \hat{g}_{k-1}) \cdot (\hat{g}_{k-1} - g_{k-1}) \cdot (\hat{\mathcal{B}}_k - \mathcal{B}_k)\right] \\ &+ P\left[m_2(\hat{\pi}) \cdot (\pi - \hat{\pi}) \cdot (\hat{\mathcal{C}}_{\mathcal{B}_k} - \mathcal{C}_{\mathcal{B}_k})\right] . \end{aligned} \tag{A.13}$$

Here, the functions m_1, m_2 and m_3 are bounded. Consequently, the overall negligibility of the second-order remainder terms depends only on the $L^2(P)$ convergence rates of the nuisance estimates in combinations corresponding to the product terms. Specifically, as long as the combined $L^2(P)$ convergence rate of the two nuisance estimates in each product term is faster than $o_p(n^{-1/2})$, the remainder term $R_2(\hat{Q}, Q)$ would also be $o_p(n^{-1/2})$. This negligibility condition enables the discussion of the asymptotic linearity of the one-step corrected plug-in estimators. Given that $\gamma^+(\hat{Q}) - \gamma(Q) = P_n(\Phi(Q)) + o_p(n^{-1/2})$, the central limit theorem implies $\sqrt{n}(\gamma^+(\hat{Q}) - \gamma) \to^d N(0, \mathbb{E}[\Phi^2(Q)])$. This is formally presented in Theorem 3.3.1.

Regarding consistency, as long as at least one component of each nuisance product term is consistently estimated (i.e., the difference between the nuisance estimate and its true value is $o_p(1)$, the one-step corrected plug-in estimator will be consistent. This robustness property is discussed in detail in Corollary 3.3.2.

B Effect decomposition

There are various ways to define path-specific effects when dealing with multiple ordered mediators, as discussed by Daniel et al. [26] and Steen et al. [74]. Assume there are K ordered mediators, $M_1, \ldots, M_k$. Generalizing (3.1) to incorporate K mediators, we can define the nested potential outcome as $Y(r_0, \mathbf{r})$, where $\mathbf{r} = (r_1, \cdots, r_K)$ with each $r_k \in \{0, 1\}$ and $r_0 \in \{0, 1\}$. The effect through M_k $(k = 1, \ldots, K)$ can be defined as a contrast of the form:

$$\tilde{\rho}_{R \to M_k \rightsquigarrow Y} = \mathbb{E}[Y(r_0, (r_1, \ldots, r_k = 1, \ldots, r_K))] - \mathbb{E}[Y(r_0, (r_1, \ldots, r_k = 0, \ldots, r_K))] \ .$$

Given the possible value combinations for r_0 and the vector $\mathbf{r}$ (with the k-th element fixed), there are 2^K potential contrasts. This also holds for the direct effect, defined as

$$\tilde{\rho}_{R \to Y} = \mathbb{E}[Y(1, \mathbf{r})] - \mathbb{E}[Y(0, \mathbf{r})] \ .$$

This flexibility allows for nuanced interpretations of how distinct pathways contribute to the overall effect, and two popular approaches to decomposing PSEs are the *sequential* and *reference-zero* decompositions. To illustrate, consider a setting with two mediators, shown in Figure B.1. Let $Y(r_0, r_1, r_2) = Y(r_0, M_1(r_1), M_2(r_2, M_1(r_1)))$ represent the potential outcome if R were set to r_0, M_1 to its natural value under $R = r_1$, and M_2 to its natural value under $R = r_2$ and $M_1(r_1)$. Below, we give examples of these two decompositions.

(1) Sequential decomposition: In this approach, specific pathways are "deactivated" in a fixed order. For the two-mediator setup shown in Figure B.1, the PSEs can be

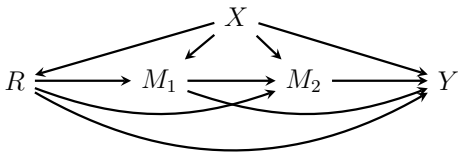


Figure B.1: A DAG with two ordered mediators.

defined as:

$$\tilde{\rho}_{R\to M_1 \rightsquigarrow Y} = \mathbb{E}[Y(1,1,1)] - \mathbb{E}[Y(1,0,1)] \ , \tag{B.1}$$

$$\tilde{\rho}_{R\to M_2 \to Y} = \mathbb{E}[Y(1,0,1)] - \mathbb{E}[Y(1,0,0)] \ , \tag{B.2}$$

$$\tilde{\rho}_{R\to Y} = \mathbb{E}[Y(1,0,0)] - \mathbb{E}[Y(0,0,0)] \ . \tag{B.3}$$

These effects are referred to as *cumulative path-specific effects* in [97]. The total effect is partitioned into $K+1$ components, with each component representing the cumulative contribution of a specific mediator to the total effect. This decomposition is particularly valuable in applications where investigators aim to quantify the proportion of the overall effect attributable to each component.

We derive the PSEs using a saturated model without confounders as an illustrative example. Consider the following expression for the mean of the nested potential outcome:

$$\begin{aligned}\mathbb{E}[Y(r_0,r_1,r_2)] =& \beta_1 r_1 + \beta_{12} r_1 r_2 + \beta_{01} r_0 r_1 + \beta_{012} r_0 r_1 r_2 \\ &+ \beta_2 r_2 + \beta_{02} r_0 r_2 \\ &+ \beta_0 r_0 \\ &+ \theta \ .\end{aligned} \tag{B.4}$$

Thus, based on (B.1) – (B.3), the PSEs are given by:

$$\tilde{\rho}_{R\to M_1 \rightsquigarrow Y} = \beta_1 + \beta_{12} + \beta_{01} + \beta_{012} \ , \quad \tilde{\rho}_{R\to M_2 \to Y} = \beta_2 + \beta_{02} \ , \quad \tilde{\rho}_{R\to Y} = \beta_0 \ .$$

Notably, $\tilde{\rho}_{R\to M_1 \rightsquigarrow Y}$ includes the main effect of r_1 (β_1) but also all interaction terms involving r_1 ($\beta_{12}, \beta_{01}, \beta_{012}$). Similarly, $\tilde{\rho}_{R\to M_2 \to Y}$ captures the main effect of r_2 (β_2) and the interaction terms involving r_2 that does not relate to r_1 (β_{02}). The direct effect, $\tilde{\rho}_{R\to Y}$, does not include any interaction terms.

(2) Reference-zero decomposition: This method focuses on specific pathways of interest, treating variables as if the treatment is set to the "active value" ($R = 1$) along the pathways of interest, while along other pathways, variables behave as if the treatment variable is set to the "baseline value" ($R = 0$). For the two-mediator setup shown in Figure B.1, the PSEs can be defined differently, as:

$$\tilde{\rho}_{R\to M_1 \rightsquigarrow Y} = \mathbb{E}[Y(0,1,0)] - \mathbb{E}[Y(0,0,0)] \ , \tag{B.5}$$

$$\tilde{\rho}_{R\to M_2 \to Y} = \mathbb{E}[Y(0,0,1)] - \mathbb{E}[Y(0,0,0)] \ , \tag{B.6}$$

$$\tilde{\rho}_{R\to Y} = \mathbb{E}[Y(1,0,0)] - \mathbb{E}[Y(0,0,0)] \ . \tag{B.7}$$

These effects are referred to as *natural path-specific effects* in [26]. Cumulative PSEs and natural PSEs share the same representation for the direct effect but differ in how they represent effects through specific mediators. Natural PSEs offer a more intuitive interpretation, such as the average change in Y if the controlled group's mediator is set to levels observed for the treatment group.

The natural PSEs derived using the model in B.4 are given by:

$$\tilde{\rho}_{R\to M_1 \rightsquigarrow Y} = \beta_1 \ , \quad \tilde{\rho}_{R\to M_2 \to Y} = \beta_2 \ , \quad \tilde{\rho}_{R\to Y} = \beta_0 \ .$$

Natural PSEs capture only the main terms $\beta_1, \beta_2, \beta_0$ (for effects through M_1, M_2, and the direct effect, respectively), excluding any interaction terms. When there are no interactions among $(r_0, \ldots, r_K)$, natural PSEs and cumulative PSEs coincide; otherwise, they can diverge—except for the direct effect, which remains the same under both definitions. Additionally, natural PSEs cannot simply be summed to obtain the total effect, nor do their proportions match the "proportion mediated" often reported in mediation analysis. Beyond these considerations, Tai et al. [77] proposed decomposing fully mediated interaction from the average causal effect, thereby offering further insight into how complex mediator interactions shape exposure–outcome relationships.

B.1 Cumulative PSEs in MEPS data

The total effect and cumulative PSEs obtained via sequential decomposition are presented as ratios of scaled geometric means in Table B.1. The product of the PSEs equals the total effect. Notably, greater deviation of a PSE from 1 indicates that the corresponding pathway accounts for a larger contribution to the racial disparities in healthcare expenditures.

Similar to the natural PSEs reported in Table 4.1, direct effects were statistically significant only in comparisons between White groups and minority groups, but not in comparisons between two minority groups in 2009. As shown by the ratios of scaled geometric means, direct effects emerged as the most dominant factor driving disparities in Whites vs minorities. This finding highlights the presence of structural discrimination in healthcare and underscores the importance of further granular investigation into unobserved factors.

Focusing specifically on the four mediators, we observe that in 2009, SES occupied as the dominant mediator in comparisons between Whites vs. Blacks and Asians vs. Hispanics. Health insurance was the primary mediator in disparities between Whites

vs. Hispanics and Blacks vs. Hispanics. Additionally, health status played the most significant role in disparities between Whites vs. Asians and Blacks vs. Asians. These findings further reinforce the conclusions presented in the main text. In particular, SES and health insurance were the most critical mediators in improving healthcare resource utilization among Hispanic individuals, emphasizing the policy relevance of expanding insurance coverage within this population. These dominant mediators mostly persisted in 2016, with the exception that health status became the most influential factor in the White vs. Black comparison. This shift may be associated with the rising prevalence of chronic diseases, potentially driven by changes in economic conditions, dietary habits, and other lifestyle factors[2].

Compared to the natural PSEs, most results were consistent across both decompositions, with a few exceptions. The effects through health status in the White vs. Black comparison, SES and health behaviors in the White vs. Asian comparison, and SES in the Black vs. Hispanic comparison showed the same direction in both the natural and cumulative PSEs, but differed in statistical significance. However, the effect of health behaviors in the White vs. Hispanic comparison reversed direction between the two decompositions. These discrepancies reflect underlying interaction effects.

Table B.1: Cumulative path-specific effects across racial group comparisons (scaled geometric mean ratios)

	MEPS data in year 2009			**MEPS data in year 2016**		
Path	Effect	95% CI	p-value	Effect	95% CI	p-value
Whites vs Blacks*						
$R \to M1 \rightsquigarrow Y$	1.161	1.125 — 1.196	**<0.001**	1.125	1.084 — 1.167	**<0.001**
$R \to M2 \rightsquigarrow Y$	0.995	0.966 — 1.025	0.745	0.997	0.971 — 1.024	0.849
$R \to M3 \rightsquigarrow Y$	0.985	0.968 — 1.003	0.096	0.997	0.978 — 1.015	0.718
$R \to M4 \to Y$	1.064	1.013 — 1.116	**0.014**	1.145	1.084 — 1.207	**<0.001**
$R \to Y$	1.764	1.609 — 1.920	**<0.001**	1.863	1.684 — 2.043	**<0.001**
Total effect	2.137	1.894 — 2.381	**<0.001**	2.387	2.106 — 2.668	**<0.001**
Whites vs Asians*						
$R \to M1 \rightsquigarrow Y$	0.887	0.859 — 0.916	**<0.001**	0.932	0.903 — 0.961	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.062	1.012 — 1.113	**0.015**	1.023	0.992 — 1.054	0.148
$R \to M3 \rightsquigarrow Y$	0.947	0.924 — 0.971	**<0.001**	0.926	0.902 — 0.950	**<0.001**
$R \to M4 \to Y$	1.323	1.237 — 1.410	**<0.001**	1.416	1.320 — 1.513	**<0.001**
$R \to Y$	2.420	2.086 — 2.755	**<0.001**	1.970	1.678 — 2.262	**<0.001**
Total effect	2.861	2.371 — 3.351	**<0.001**	2.462	2.047 — 2.876	**<0.001**
Whites vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.282	1.202 — 1.362	**<0.001**	1.252	1.176 — 1.327	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.409	1.338 — 1.480	**<0.001**	1.417	1.351 — 1.483	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.911	0.856 — 0.967	**0.002**	0.912	0.859 — 0.966	**0.001**
$R \to M4 \to Y$	1.332	1.237 — 1.427	**<0.001**	1.393	1.287 — 1.499	**<0.001**
$R \to Y$	2.113	1.931 — 2.296	**<0.001**	1.907	1.739 — 2.075	**<0.001**
Total effect	4.633	4.141 — 5.126	**<0.001**	4.298	3.819 — 4.776	**<0.001**
Blacks vs Asians*						
$R \to M1 \rightsquigarrow Y$	0.820	0.743 — 0.897	**<0.001**	0.738	0.669 — 0.807	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.064	1.003 — 1.125	**0.038**	1.012	0.969 — 1.055	0.587
$R \to M3 \rightsquigarrow Y$	1.009	0.976 — 1.041	0.604	0.995	0.960 — 1.030	0.798
$R \to M4 \to Y$	1.426	1.267 — 1.585	**<0.001**	1.498	1.337 — 1.659	**<0.001**
$R \to Y$	1.045	0.876 — 1.214	0.602	0.881	0.745 — 1.016	0.083
Total effect	1.311	1.033 — 1.589	**0.028**	0.981	0.783 — 1.178	0.848
Blacks vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.086	0.979 — 1.194	0.116	1.119	1.031 — 1.206	**0.008**
$R \to M2 \rightsquigarrow Y$	1.449	1.318 — 1.579	**<0.001**	1.414	1.327 — 1.500	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.958	0.894 — 1.021	0.193	0.958	0.906 — 1.011	0.122
$R \to M4 \to Y$	1.356	1.211 — 1.502	**<0.001**	1.282	1.164 — 1.400	**<0.001**
$R \to Y$	1.023	0.943 — 1.103	0.577	0.875	0.793 — 0.957	**0.003**
Total effect	2.091	1.779 — 2.403	**<0.001**	1.701	1.457 — 1.945	**<0.001**
Asians vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	1.868	1.579 — 2.157	**<0.001**	1.727	1.442 — 2.011	**<0.001**
$R \to M2 \rightsquigarrow Y$	1.219	1.119 — 1.320	**<0.001**	1.222	1.058 — 1.387	**0.008**
$R \to M3 \rightsquigarrow Y$	0.977	0.930 — 1.024	0.340	0.975	0.915 — 1.035	0.413
$R \to M4 \to Y$	0.819	0.703 — 0.935	**0.002**	0.773	0.587 — 0.959	**0.017**
$R \to Y$	1.019	0.943 — 1.095	0.624	1.169	1.070 — 1.268	**0.001**
Total effect	1.858	1.523 — 2.194	**<0.001**	1.860	1.534 — 2.187	**<0.001**

*Reference group; M_1: SES, M_2: Insurance, M_3: Health behaviors, M_4: Health status.

C The responses in MEPS data

C.1 Geometric mean interpretation

Positive responses

Assume responses are all positive. Consider the potential outcome $Y(r_0, \mathbf{r})$, defined in (3.1). Let $Y(r_0, \mathbf{r}) = \log Y(r_0, \mathbf{r})$. The effects are defined as:

$$\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] = \mathbb{E}\left[\log Y(r_0, \mathbf{r}) - \log Y(0, \mathbf{0})\right] = \mathbb{E}\left[\log \frac{Y(r_0, \mathbf{r})}{Y(0, \mathbf{0})}\right] .$$

To interpret the above on the scale similar to the healthcare expenditures, we exponentiate $\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$:

$$\begin{aligned} \exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]\right) &= \exp\left(\mathbb{E}\left[\log \frac{Y(r_0, \mathbf{r})}{Y(0, \mathbf{0})}\right]\right) \\ &\approx \frac{\left\{\prod_{i=1}^n Y_i(r_0, \mathbf{r})\right\}^{1/n}}{\left\{\prod_{i=1}^n Y_i(0, \mathbf{0})\right\}^{1/n}} = \frac{G_n\big(Y(r_0, \mathbf{r})\big)}{G_n\big(Y(0, \mathbf{0})\big)} , \end{aligned}$$

where $G_n(f)$ denotes the geometric mean of f, i.e., $G_n(f) = \{\prod_{i=1}^n f_i\}^{1/n}$.

We note that identification and estimation arguments for $\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$ remain the same by simply defining the outcome as log of healthcare expenditures. The identification functionals are given by:

$$\begin{aligned} \mathbb{E}[Y(0, \mathbf{0})] &= \int \log y \times dP(y \mid R = 0, x) \times dP(x) , \\ \mathbb{E}[Y(1, \mathbf{0})] &= \int \log y \times dP(y \mid \overline{m}_4, R = 1, x) \times \prod_{k=1}^{4} dP(m_k \mid \overline{m}_{k-1}, R = 0, x) \times dP(x) , \\ \mathbb{E}[Y(0, \mathbf{1}_k)] &= \int \log y \times dP(y \mid \overline{m}_4, R = 0, x) \times dP(m_k \mid \overline{m}_{k-1}, R = 1, x) \times \end{aligned}$$

$$\prod_{\substack{j=1 \\ j\neq k}}^{4} dP(m_j \mid \overline{m}_{j-1}, R=0, x) \times dP(x) \ . \tag{C.1}$$

Positive and zero responses

In our setting, we have both positive and zero responses. Let $Y(r_0, \mathbf{r}) = \mathbb{I}(Y(r_0, \mathbf{r}) > 0) \log Y(r_0, \mathbf{r})$. The effects are defined as:

$$\begin{aligned}
\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] &= \mathbb{E}\left[\mathbb{I}(Y(r_0, \mathbf{r}) > 0) \log Y(r_0, \mathbf{r}) - \mathbb{I}(Y(0, \mathbf{0}) > 0) \log Y(0, \mathbf{0})\right] \\
&= P(Y(r_0, \mathbf{r}) > 0) \times \mathbb{E}\left[\log Y(r_0, \mathbf{r}) \mid Y(r_0, \mathbf{r}) > 0\right] \\
&\quad - P(Y(0, \mathbf{0}) > 0) \times \mathbb{E}\left[\log Y(0, \mathbf{0}) \mid Y(0, \mathbf{0}) > 0\right] \ .
\end{aligned}$$

To interpret the above on the scale similar to the healthcare expenditures, we exponentiate $\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$:

$$\begin{aligned}
\exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]\right) &= \frac{\exp\left(P(Y(r_0, \mathbf{r}) > 0) \times \mathbb{E}\left[\log Y(r_0, \mathbf{r}) \mid Y(r_0, \mathbf{r}) > 0\right]\right)}{\exp\left(P(Y(0, \mathbf{0}) > 0) \times \mathbb{E}\left[\log Y(0, \mathbf{0}) \mid Y(0, \mathbf{0}) > 0\right]\right)} \\
&= \frac{\left\{ \exp\left(\mathbb{E}\left[\log Y(r_0, \mathbf{r}) \mid Y(r_0, \mathbf{r}) > 0\right]\right)\right\}^{P(Y(r_0, \mathbf{r})>0)}}{\left\{ \exp\left(\mathbb{E}\left[\log Y(0, \mathbf{0}) \mid Y(0, \mathbf{0}) > 0\right]\right)\right\}^{P(Y(0, \mathbf{0})>0)}} \\
&\approx \frac{\left\{ \prod_{i=1}^{n} Y_{i,\text{pos}}(r_0, \mathbf{r})\right\}^{\hat{P}(Y(r_0, \mathbf{r})>0)/n}}{\left\{ \prod_{i=1}^{n} Y_{i,\text{pos}}(0, \mathbf{0})\right\}^{\hat{P}(Y(0, \mathbf{0})>0)/n}} \\
&= \frac{G_n\big(Y_{\text{pos}}(r_0, \mathbf{r})\big)^{\hat{P}(Y(r_0, \mathbf{r})>0)}}{G_n\big(Y_{\text{pos}}(0, \mathbf{0})\big)^{\hat{P}(Y(0, \mathbf{0})>0)}} \ ,
\end{aligned} \tag{C.2}$$

where $G_n(Y_{\text{pos}}(r_0, \mathbf{r}))$ and $G_n(Y_{\text{pos}}(0, \mathbf{0}))$ denote the geometric mean of positive counterfactual responses $Y_{\text{pos}}(r_0, \mathbf{r})$ and $G_n(Y_{\text{pos}}(0, \mathbf{0}))$, respectively. Therefore, the effect can be interpreted as ratio of scaled geometric means.

We note that identification and estimation arguments for $\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$ re-

main the same by simply defining the outcome as zero if expenditure is zero, and log of expenditure otherwise. The identification functionals are given by:

$$\begin{aligned}
\mathbb{E}[Y(0,\mathbf{0})] &= \int \mathbb{I}(y>0)\log y \times dP(y \mid R=0,x) \times dP(x)\ , \\
\mathbb{E}[Y(1,\mathbf{0})] &= \int \mathbb{I}(y>0)\log y \times dP(y \mid \overline{m}_4, R=1,x) \times \prod_{k=1}^{4} dP(m_k \mid \overline{m}_{k-1}, R=0,x) \times dP(x)\ , \\
\mathbb{E}[Y(0,\mathbf{1}_k)] &= \int \mathbb{I}(y>0)\log y \times dP(y \mid \overline{m}_4, R=0,x) \times dP(m_k \mid \overline{m}_{k-1}, R=1,x) \times \\
&\qquad \prod_{\substack{j=1 \\ j\neq k}}^{4} dP(m_j \mid \overline{m}_{j-1}, R=0,x) \times dP(x)\ . \qquad \text{(C.3)}
\end{aligned}$$

Remark 1 (**Asymptotic variance)**. By delta method, we can write:

$$\begin{aligned}
&\sqrt{n}(\exp(\rho^{+}_{R\to Y}(\hat{Q})) - \exp(\rho_{R\to Y}(Q))) \\
&\qquad \to^{d} \mathcal{N}\left(0, \exp(\rho_{R\to Y}(Q))^2 \times \mathbb{E}[\left(\Phi_{\gamma_{R\to Y}}(Q) - \Phi_{\gamma_{\text{inact}}}(Q)\right)^2]\right)\ ,
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{n}(\exp(\rho^{+}_{R\to M_k \rightsquigarrow Y}(\hat{Q})) - \exp(\rho_{R\to M_k \rightsquigarrow Y}(Q))) \\
&\qquad \to^{d} \mathcal{N}\left(0, \exp(\rho_{R\to M_k \rightsquigarrow Y}(Q))^2 \times \mathbb{E}[\left(\Phi_{\gamma_{R\to M_k \rightsquigarrow Y}}(Q) - \Phi_{\gamma_{\text{inact}}}(Q)\right)^2]\right)\ .
\end{aligned}$$

Remark 2 (**Probability of positive counterfactual responses)**. In addition to reporting effects with the interpretations outlined in (C.2), we also report effects based on a binary indicator for zero or positive responses in table C.1, i.e., $P(Y(r_0,\mathbf{r}) > 0) - P(Y(0,\mathbf{0}) > 0)$. The identification and estimation arguments remain unchanged, with the outcome simply redefined as $\mathbb{I}(Y>0)$.

Remark 3 (**Smearing transformation)**. The smearing transformation is often applied to adjust for the bias introduced when exponentiating $\mathbb{E}[Y(r_0,\mathbf{r}) - Y(0,\mathbf{0})]$ to

estimate the arithmetic mean of the differences, $\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$, rather than the geometric mean. As an example, assume:

$$
\begin{aligned}
Y(r_0, \mathbf{r}) - Y(0, \mathbf{0}) &\sim \mathcal{N}(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})], \sigma^2) \\
Y(r_0, \mathbf{r}) - Y(0, \mathbf{0}) &= \mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] + \epsilon_i, \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \ .
\end{aligned}
$$

Therefore:

$$
Y(r_0, \mathbf{r}) - Y(0, \mathbf{0}) = \exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] + \epsilon\right) \ ,
$$

and

$$
\begin{aligned}
\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] &= \mathbb{E}\left[\exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})] + \epsilon\right)\right] \\
&= \exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]\right) \times \mathbb{E}\left[\exp\left(\epsilon\right)\right] \\
&= \exp\left(\mathbb{E}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]\right) \times \exp\left(\sigma^2/2\right) \ .
\end{aligned}
$$

The last equality holds by the moment-generating function of a Normal distribution. Here, σ^2 is the variance of $Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})$, that is the variance of the difference between the log-transformed $Y(r_0, \mathbf{r})$ and log-transformed $Y(0, \mathbf{0})$.

If the assumption of a normally distributed error term is violated, the empirical mean can be used to estimate $\mathbb{E}\left[\exp\left(\epsilon\right)\right]$, specifically as $\frac{1}{n}\sum_{i=1}^{n}\exp(\epsilon_i)$, where $\epsilon_i = Y_i(r_0, \mathbf{r}) - Y_i(0, \mathbf{0}) - \hat{\mathbb{E}}[Y(r_0, \mathbf{r}) - Y(0, \mathbf{0})]$.

C.2 Two-stage super learner

Let $Y(r_0, \mathbf{r}) = Y(r_0, \mathbf{r})$, where the outcome is defined as the original healthcare expenditures, which include both positive and zero responses. The effects, as defined in 3.3,

are interpreted as differences in arithmetic means. To obtain the one-step estimates, outlined in 3.9, the function $\mu_k(\overline{M}_k, r_0, X)$ was estimated using the two-stage super learner, as demonstrated in an example here [link]. The two-stage super learner library comprises all pairwise combinations of two constituent algorithms: one for estimating $P(Y > 0 \mid \overline{M}_k, r_0, X)$ and another for $\mathbb{E}[Y \mid Y > 0, \overline{M}_k, r_0, X]$. Using a two-stage super learner is expected to improve predictions for each individual outcome.

Table C.2 presents the results of PSEs calculated as differences in arithmetic means. These findings differ notably from those in Table 4.1 and Table C.1, where results in the latter two tables are mostly aligned. For instance, the effect through SES ($R \to M_1 \rightsquigarrow Y$) for Whites vs. Blacks and the total effect for Asians vs. Hispanics were significantly positive in Table 4.1 and Table C.1 but became significantly negative in Table C.2. These discrepancies underscore the risks of directly using arithmetic means in the analysis of skewed data, which may lead to potential misinterpretations of the results.

Table C.1: Path-specific effects for different racial group comparisons on the probability of positive healthcare expenditures, reported on the difference scale.

	MEPS data in year 2009			**MEPS data in year 2016**		
Path	Effect	95% CI	p value	Effect	95% CI	p value
Whites vs Blacks*						
$R \to M1 \rightsquigarrow Y$	0.016	0.010 — 0.023	**<0.001**	0.024	0.018 — 0.031	**<0.001**
$R \to M2 \rightsquigarrow Y$	0.001	-0.003 — 0.005	0.628	0.001	-0.002 — 0.004	0.495
$R \to M3 \rightsquigarrow Y$	-0.001	-0.004 — 0.002	0.516	0.000	-0.002 — 0.002	0.779
$R \to M4 \to Y$	0.000	-0.007 — 0.007	0.953	0.009	0.002 — 0.017	**0.012**
$R \to Y$	0.057	0.045 — 0.069	**<0.001**	0.061	0.048 — 0.074	**<0.001**
Total effect	0.075	0.061 — 0.089	**<0.001**	0.091	0.077 — 0.104	**<0.001**
Whites vs Asians*						
$R \to M1 \rightsquigarrow Y$	-0.010	-0.021 — 0.002	0.117	-0.010	-0.019 — -0.001	**0.034**
$R \to M2 \rightsquigarrow Y$	0.009	0.001 — 0.017	**0.023**	0.002	-0.003 — 0.006	0.424
$R \to M3 \rightsquigarrow Y$	-0.003	-0.008 — 0.002	0.236	-0.002	-0.006 — 0.002	0.323
$R \to M4 \to Y$	0.025	0.009 — 0.040	**0.002**	0.024	0.011 — 0.037	**<0.001**
$R \to Y$	0.063	0.043 — 0.083	**<0.001**	0.055	0.037 — 0.074	**<0.001**
Total effect	0.069	0.047 — 0.092	**<0.001**	0.063	0.043 — 0.083	**<0.001**
Whites vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	0.048	0.038 — 0.058	**<0.001**	0.047	0.038 — 0.057	**<0.001**
$R \to M2 \rightsquigarrow Y$	0.036	0.030 — 0.042	**<0.001**	0.037	0.031 — 0.042	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.006	-0.001 — 0.014	0.090	0.003	-0.003 — 0.010	0.335
$R \to M4 \to Y$	0.031	0.022 — 0.039	**<0.001**	0.043	0.034 — 0.051	**<0.001**
$R \to Y$	0.084	0.072 — 0.096	**<0.001**	0.069	0.057 — 0.081	**<0.001**
Total effect	0.163	0.150 — 0.177	**<0.001**	0.148	0.135 — 0.161	**<0.001**
Blacks vs Asians*						
$R \to M1 \rightsquigarrow Y$	-0.016	-0.038 — 0.006	0.147	-0.028	-0.047 — -0.009	**0.003**
$R \to M2 \rightsquigarrow Y$	0.009	0.000 — 0.017	**0.048**	0.002	-0.004 — 0.007	0.530
$R \to M3 \rightsquigarrow Y$	-0.005	-0.010 — 0.001	0.122	-0.002	-0.007 — 0.003	0.407
$R \to M4 \to Y$	0.030	0.011 — 0.048	**0.002**	0.023	0.008 — 0.037	**0.003**
$R \to Y$	-0.019	-0.043 — 0.005	0.124	-0.026	-0.048 — -0.004	**0.020**
Total effect	-0.010	-0.038 — 0.019	0.515	-0.038	-0.063 — -0.013	**0.003**
Blacks vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	0.023	0.016 — 0.030	**<0.001**	0.014	0.008 — 0.020	**<0.001**
$R \to M2 \rightsquigarrow Y$	0.045	0.037 — 0.052	**<0.001**	0.039	0.033 — 0.045	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.004	-0.001 — 0.009	0.163	0.002	-0.004 — 0.008	0.499
$R \to M4 \to Y$	0.031	0.022 — 0.041	**<0.001**	0.022	0.014 — 0.031	**<0.001**
$R \to Y$	0.007	-0.005 — 0.019	0.253	-0.016	-0.029 — -0.004	**0.010**
Total effect	0.088	0.068 — 0.108	**<0.001**	0.056	0.037 — 0.075	**<0.001**
Asians vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	0.075	0.060 — 0.089	**<0.001**	0.068	0.055 — 0.081	**<0.001**
$R \to M2 \rightsquigarrow Y$	0.028	0.018 — 0.038	**<0.001**	0.033	0.024 — 0.042	**<0.001**
$R \to M3 \rightsquigarrow Y$	0.000	-0.002 — 0.003	0.900	0.001	-0.001 — 0.004	0.287
$R \to M4 \to Y$	-0.012	-0.025 — 0.001	0.062	-0.013	-0.027 — 0.000	0.058
$R \to Y$	0.029	0.018 — 0.041	**<0.001**	0.032	0.021 — 0.043	**<0.001**
Total effect	0.111	0.086 — 0.136	**<0.001**	0.099	0.076 — 0.123	**<0.001**

* Reference group; M_1: SES, M_2: Insurance access, M_3: Health behaviors, M_4: Health status.

Table C.2: Path-specific effects for different racial group comparisons using two-stage super learner, reported on the difference scale (arithmetic mean).

	MEPS data in year 2009			**MEPS data in year 2016**		
Path	Effect	95% CI	p value	Effect	95% CI	p value
Whites vs Blacks*						
$R \to M1 \rightsquigarrow Y$	-167.6	-448.6 — 113.4	0.242	-129.0	-406.3 — 148.2	0.362
$R \to M2 \rightsquigarrow Y$	-18.1	-80.0 — 43.9	0.567	-17.0	-65.2 — 31.3	0.491
$R \to M3 \rightsquigarrow Y$	-77.7	-191.2 — 35.7	0.179	27.2	-57.5 — 111.8	0.529
$R \to M4 \to Y$	388.3	11.7 — 764.9	**0.043**	757.1	349.0 — 1165.3	**<0.001**
$R \to Y$	521.3	7.4 — 1035.2	**0.047**	1,322.8	748.0 — 1897.6	**<0.001**
Total effect	161.3	-353.9 — 676.5	0.540	1,022.2	407.7 — 1636.7	**0.001**
Whites vs Asians*						
$R \to M1 \rightsquigarrow Y$	6.6	-206.6 — 219.8	0.952	291.0	-677.8 — 1259.8	0.556
$R \to M2 \rightsquigarrow Y$	109.3	38.6 — 180.1	**0.002**	32.5	-64.7 — 129.6	0.512
$R \to M3 \rightsquigarrow Y$	-30.7	-118.3 — 56.9	0.492	89.3	-2.5 — 181.0	0.056
$R \to M4 \to Y$	1,167.4	802.1 — 1532.7	**<0.001**	1,666.1	1082.3 — 2250.0	**<0.001**
$R \to Y$	1,973.5	1562.8 — 2384.2	**<0.001**	1,773.4	1088.2 — 2458.6	**<0.001**
Total effect	2,512.2	2032.7 — 2991.7	**<0.001**	2,834.0	2122.2 — 3545.7	**<0.001**
Whites vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	-90.7	-336.9 — 155.5	0.470	376.1	-14.8 — 767.0	0.059
$R \to M2 \rightsquigarrow Y$	432.8	331.7 — 533.9	**<0.001**	377.5	294.4 — 460.6	**<0.001**
$R \to M3 \rightsquigarrow Y$	80.1	-38.4 — 198.5	0.185	159.8	-26.3 — 345.8	0.092
$R \to M4 \to Y$	1,451.9	1143.7 — 1760.1	**<0.001**	1,712.7	1389.7 — 2035.8	**<0.001**
$R \to Y$	787.1	452.4 — 1121.9	**<0.001**	1,148.7	740.8 — 1556.7	**<0.001**
Total effect	1,543.1	1194.7 — 1891.5	**<0.001**	2,115.8	1626.0 — 2605.7	**<0.001**
Blacks vs Asians*						
$R \to M1 \rightsquigarrow Y$	53.9	-207.9 — 315.6	0.687	329.5	-106.7 — 765.7	0.139
$R \to M2 \rightsquigarrow Y$	40.5	-45.8 — 126.7	0.358	17.2	-77.1 — 111.6	0.720
$R \to M3 \rightsquigarrow Y$	-44.3	-140.8 — 52.2	0.368	89.6	-36.6 — 215.8	0.164
$R \to M4 \to Y$	1,682.3	1204.6 — 2160.1	**<0.001**	2,139.4	1606.9 — 2671.9	**<0.001**
$R \to Y$	1,087.0	662.2 — 1511.8	**<0.001**	650.8	117.0 — 1184.7	**0.017**
Total effect	2,176.0	1628.4 — 2723.7	**<0.001**	1,695.2	1031.1 — 2359.2	**<0.001**
Blacks vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	25.3	-120.3 — 171.0	0.733	284.7	120.9 — 448.5	**0.001**
$R \to M2 \rightsquigarrow Y$	526.9	400.7 — 653.1	**<0.001**	406.8	323.0 — 490.7	**<0.001**
$R \to M3 \rightsquigarrow Y$	60.5	-25.2 — 146.3	0.166	46.5	-102.3 — 195.3	0.541
$R \to M4 \to Y$	1,242.7	914.6 — 1570.8	**<0.001**	954.8	594.7 — 1314.8	**<0.001**
$R \to Y$	249.0	-49.7 — 547.6	0.102	104.2	-253.2 — 461.6	0.568
Total effect	1,146.0	683.3 — 1608.6	**<0.001**	854.6	305.4 — 1403.8	**0.002**
Asians vs Hispanics*						
$R \to M1 \rightsquigarrow Y$	84.5	-275.8 — 444.8	0.646	553.8	229.3 — 878.3	**0.001**
$R \to M2 \rightsquigarrow Y$	298.6	169.6 — 427.6	**<0.001**	258.7	148.9 — 368.4	**<0.001**
$R \to M3 \rightsquigarrow Y$	-9.6	-71.7 — 52.5	0.762	26.8	-38.4 — 92.1	0.420
$R \to M4 \to Y$	-391.8	-697.9 — -85.6	**0.012**	-527.4	-917.5 — -137.3	**0.008**
$R \to Y$	-50.0	-319.0 — 219.1	0.716	370.0	42.6 — 697.4	**0.027**
Total effect	-719.2	-1089.0 — -349.3	**<0.001**	-596.1	-1062.3 — -130.0	**0.012**

* Reference group; M_1: SES, M_2: Insurance access, M_3: Health behaviors, M_4: Health status.

D Additional simulation

The variables $(X_1, X_2, X_3, R, M_{11}, M_{12}, M_2, M_{31}, M_{32}, M_{41}, M_{42}, Y)$ in this simulation study are generated via the following models:

$$
\begin{aligned}
&X_1, X_2 \overset{iid}{\sim} \text{Uniform}(0,1), \quad X_3 \sim \text{Bernoulli}(0.5), \quad R \sim \text{Bernoulli}(\text{expit}(V_R \begin{bmatrix} 1 & X_1 & X_2 & X_3 \end{bmatrix}^T)),\\
&M_1 = \begin{bmatrix} M_{11} & M_{12} \end{bmatrix}, M_{12} \sim \text{Bernoulli}(\text{expit}(M_{12}^*)),\\
&\qquad \begin{bmatrix} M_{11} \\ M_{12}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} V_{M_{11}}(1 & R & X_1X_2 & X_3)^T \\ V_{M_{11}}(1 & R & X_1^2 & X_2 * X_3)^T \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),\\
&M_2 \sim \text{Bernoulli}(\text{expit}(V_{M_2} \begin{bmatrix} 1 & -RX_3 & M_1 & X_1 & X_2 \end{bmatrix}^T)),\\
&M_3 = \begin{bmatrix} M_{31} & M_{32} \end{bmatrix}, M_{31} \sim \text{Bernoulli}(\text{expit}(M_{31}^*)), M_{32} \sim \text{Bernoulli}(\text{expit}(M_{32}^*)),\\
&\qquad \begin{bmatrix} M_{31}^* \\ M_{32}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} V_{M_{31}}(1 & R & M_1 & M_2 & X_1 & X_2 & RX_3)^T \\ V_{M_{32}}(1 & R & M_1 & M_2 & X_1^{0.5} & X_2 & X_3)^T \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \right),\\
&M_4 = \begin{bmatrix} M_{41} & M_{42} \end{bmatrix}, M_{42} \sim \text{Bernoulli}(\text{expit}(M_{42}^*)),\\
&\qquad \begin{bmatrix} M_{41} \\ M_{42}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} V_{M_{41}}(1 & R & M_1 & M_2 & M_3 & X_1 & X_2 & X_2X_3)^T \\ V_{M_{42}}(1 & R & M_1 & M_2 & M_3 & X_1 & X_2^2 & X_3)^T \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right),\\
&Y^* = V_Y \begin{bmatrix} 1 & R & M_1 & M_2 & M_3 & M_4 & M_{41}X_1 & X_2 & RX_3 \end{bmatrix}^T,\\
&\mathbb{I}(Y > 0) \sim \text{Bernoulli}(\text{expit}(Y^*)),\\
&Y \mid Y > 0 \sim \text{LogNormal}(\log\mu = 0.5Y^*, \log sd = 0.5).
\end{aligned} \tag{D.1}
$$

where

$$V_R = [-0.12, 0.12, 0.24, -0.2],$$

$$V_{M_{11}} = [0.33, 0.04, 0.25, 0.35],$$

$$V_{M_{12}} = [0.48, 0.4, 0.49, 0.23],$$

$$V_{M_2} = [0.14, 0.02, -0.21, 0.05, -0.4, -0.16],$$

$$V_{M_{31}} = [0.42, -0.01, -0.02, -0.05, 0.43, -0.07, -0.2, -0.18],$$

$$V_{M_{32}} = [0.36, 0.38, 0.1, -0.13, -0.05, 0.39, 0.18, -0.1],$$

$$V_{M_{41}} = [-0.01, 0.49, 0.29, -0.07, 0.41, 0.01, -0.1, -0.1, 0.13, 0.38],$$

$$V_{M_{42}} = [-0.15, 0.18, 0.4, -0.13, -0.15, 0.21, 0.07, 0.38, 0.38, 0.44],$$

$$V_Y = [0.14, 0.29, -0.44, 1, 0.47, 0.09, 0.31, 0.88, 0.34, 0.81, 0.92, 0.98].$$

We adopt the same variable generation strategy as described in Chapter 5.2, but with a simplified data structure that more closely resembles MEPS. Specifically, we use only three covariates, with M_2 as a unidimensional binary variable and M_3 as a two-dimensional variable, where each dimension is binary.

Table D.1: Comparative performance of one-step estimator using super learner (SL) vs. GLM in MEPS data structure

	Bias		SD		MSE		Coverage Rate		CI width	
sample size	SL	GLM	SL	GLM	SL	GLM	SL	GLM	SL	GLM
$\rho^+_{R\to M_1\rightsquigarrow Y}$										
1000	0.001	0.000	0.016	0.015	0.000	0.000	0.943	0.954	0.061	0.058
2000	0.001	0.000	0.011	0.010	0.000	0.000	0.930	0.949	0.040	0.040
4000	0.000	0.000	0.007	0.007	0.000	0.000	0.936	0.950	0.027	0.028
8000	0.000	0.000	0.005	0.005	0.000	0.000	0.936	0.955	0.019	0.020
$\rho^+_{R\to M_2\rightsquigarrow Y}$										
1000	0.002	0.001	0.017	0.016	0.000	0.000	0.924	0.959	0.057	0.061
2000	0.000	0.000	0.011	0.011	0.000	0.000	0.929	0.953	0.039	0.043
4000	0.000	0.000	0.008	0.008	0.000	0.000	0.934	0.950	0.028	0.030
8000	0.000	0.000	0.005	0.005	0.000	0.000	0.951	0.960	0.020	0.021
$\rho^+_{R\to M_3\rightsquigarrow Y}$										
1000	0.001	0.001	0.009	0.008	0.000	0.000	0.942	0.963	0.034	0.030
2000	0.000	0.000	0.005	0.005	0.000	0.000	0.930	0.954	0.019	0.019
4000	0.000	0.000	0.003	0.003	0.000	0.000	0.926	0.956	0.012	0.013
8000	0.000	0.000	0.002	0.002	0.000	0.000	0.925	0.947	0.008	0.009
$\rho^+_{R\to M_4\to Y}$										
1000	-0.006	-0.001	0.060	0.061	0.004	0.004	0.900	0.939	0.207	0.234
2000	-0.001	0.003	0.040	0.041	0.002	0.002	0.938	0.955	0.151	0.165
4000	-0.001	0.001	0.030	0.030	0.001	0.001	0.926	0.946	0.109	0.117
8000	-0.001	0.000	0.021	0.021	0.000	0.000	0.933	0.941	0.078	0.082
$\rho^+_{R\to Y}$										
1000	-0.006	0.000	0.056	0.057	0.003	0.003	0.904	0.947	0.192	0.221
2000	0.000	0.003	0.038	0.039	0.001	0.002	0.931	0.948	0.140	0.156
4000	0.000	0.002	0.027	0.027	0.001	0.001	0.948	0.963	0.101	0.110
8000	-0.001	0.000	0.019	0.019	0.000	0.000	0.948	0.959	0.072	0.078

Bibliography

[1] Ruopeng An. Health care expenses in relation to obesity and smoking among US adults by gender, race/ethnicity, and age group: 1998–2011. *Public Health*, 129 (1):29–36, 2015.

[2] John P Ansah and Chi-Tsun Chiu. Projecting the chronic disease burden among the adult population in the united states using a multi-state population model. *Frontiers in Public Health*, 10:1082183, 2023.

[3] Zinzi D Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T Bassett. Structural racism and health inequities in the USA: evidence and interventions. *The Lancet*, 389(10077):1453–1463, 2017.

[4] Zinzi D. Bailey, Justin M. Feldman, and Mary T. Bassett. How structural racism works — racist policies as a root cause of U.S. racial health inequities. *New England Journal of Medicine*, 384(8):768–773, 2021. doi: 10.1056/NEJMms2025396.

[5] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[6] Edwine Barasa, Peter Nguhiu, and Di McIntyre. Measuring progress towards sustainable development goal 3.8 on universal health coverage in kenya. *BMJ Global Health*, 3(3):e000904, 2018.

[7] Geoffrey S Barkley. Factors influencing health behaviors in the national health and nutritional examination survey, III (NHANES III). *Social Work in Health Care*, 46(4):57–79, 2008.

[8] Reuben M Baron and David A Kenny. The moderator–mediator variable dis-

tinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

[9] Caryn N Bell, Roland J Thorpe, and Thomas A LaVeist. The role of social context in racial disparities in self-rated health. *Journal of Urban Health*, 95:13–20, 2018.

[10] Federico Belotti, Partha Deb, Willard G Manning, and Edward C Norton. twopm: Two-part models. *The Stata Journal*, 15(1):3–20, 2015.

[11] David Benkeser and Jialu Ran. Nonparametric inference for interventional effects with multiple mediators. *Journal of Causal Inference*, 9(1):172–189, 2021.

[12] MA Beydoun, HA Beydoun, N Mode, GA Dore, JA Canas, SM Eid, and AB Zonderman. Racial disparities in adult all-cause and cause-specific mortality among US adults: mediating and moderating factors. *BMC Public Health*, 16:1–13, 2016.

[13] Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[14] Aaron J Boulton and Anne Williford. Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers. *Journal of the Society for Social Work and Research*, 9(4):721–740, 2018.

[15] Paula A Braveman, Shiriki Kumanyika, Jonathan Fielding, Thomas LaVeist, Luisa N Borrell, Ron Manderscheid, and Adewale Troutman. Health disparities and health equity: the issue is justice. *American Journal of Public Health*, 101 (S1):S149–S155, 2011.

[16] Thomas C Buchmueller and Helen G Levy. The ACA's impact on racial and ethnic disparities in health insurance coverage and access to care: an examination

of how the insurance coverage expansions of the Affordable Care Act have affected disparities related to race and ethnicity. *Health Affairs*, 39(3):395–402, 2020.

[17] Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5:892837, 2022.

[18] Centers for Disease Control and Prevention. Current cigarette smoking among adults in the United States, October 2023. URL `https://www.cdc.gov/tobacco/php/data-statistics/adult-data-cigarettes/index.html`.

[19] Centers for Disease Control and Prevention. Socioeconomic factors, September 2023. URL `https://www.cdc.gov/dhdsp/health_equity/socioeconomic.htm`.

[20] Centers for Disease Control and Prevention. Adult physical inactivity outside of work, May 2024. URL `https://www.cdc.gov/physical-activity/php/data/inactivity-maps.html`.

[21] Raphaël Charron-Chénier and Collin W Mueller. Racial disparities in medical spending: healthcare expenditures for black and white households (2013–2015). *Race and Social Problems*, 10:113–133, 2018.

[22] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1): C1–C68, 2018.

[23] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the Thirty Third Conference on Association for the Advancement of Artificial Intelligence (AAAI-33rd)*. AAAI Press, 2019.

[24] Benjamin Lê Cook and Willard G Manning. Measuring racial/ethnic disparities across the distribution of health care expenditures. *Health Services Research*, 44 (5p1):1603–1621, 2009.

[25] Lisa A Cooper, Debra L Roter, Kathryn A Carson, Mary Catherine Beach, Janice A Sabin, Anthony G Greenwald, and Thomas S Inui. The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *American Journal of Public Health*, 102(5): 979–987, 2012.

[26] Rhian M Daniel, Bianca L De Stavola, Simon N Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

[27] Karen Davis. Achievements and problems of medicaid. *Public Health Reports*, 91 (4):309, 1976.

[28] Iván Díaz. Non-agency interventions for causal mediation in the presence of intermediate confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):435–460, 2024.

[29] Iván Díaz and Nima S Hejazi. Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):661–683, 2020.

[30] Iván Díaz, Nima S Hejazi, Kara E Rudolph, and Mark J van Der Laan. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*, 108 (3):627–641, 2021.

[31] Samuel L Dickman, Adam Gaffney, Alecia McGregor, David U Himmelstein, Danny McCormick, David H Bor, and Steffie Woolhandler. Trends in health

care use among black and white persons in the US, 1963-2019. *JAMA Network Open*, 5(6):e2217383–e2217383, 2022.

[32] Joseph L Dieleman, Carina Chen, Sawyer W Crosby, Angela Liu, Darrah McCracken, Ian A Pollock, Maitreyi Sahu, Golsum Tsakalos, Laura Dwyer-Lindgren, Annie Haakenstad, et al. US health care spending by race and ethnicity, 2002-2016. *Jama*, 326(7):649–659, 2021.

[33] Chyke A Doubeni, Douglas A Corley, Wei Zhao, YanKwan Lau, Christopher D Jensen, and Theodore R Levin. Association between improved colorectal screening and racial disparities. *New England Journal of Medicine*, 386(8):796–798, 2022.

[34] Ivan Frankovic and Michael Kuhn. Health insurance, endogenous medical progress, health expenditure growth, and welfare. *Journal of Health Economics*, 87:102717, 2023.

[35] Adam Gaffney and Danny McCormick. The Affordable Care Act: implications for health-care equity. *The Lancet*, 389(10077):1442–1452, 2017.

[36] Clark Glymour and Madelyn R Glymour. Commentary: race and sex are causes. *Epidemiology*, 25(4):488–490, 2014.

[37] Enkai Guo, Huamei Zhong, Yang Gao, Jing Li, and Zhaohong Wang. Socioeconomic disparities in health care consumption: using the 2018-China family panel studies. *International Journal of Environmental Research and Public Health*, 19 (12):7359, 2022.

[38] William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. Implicit racial/ethnic bias among health care professionals and its influence on

health care outcomes: a systematic review. *American Journal of Public Health*, 105(12):e60–e76, 2015.

[39] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, 2020.

[40] Latoya Hill, Samantha Artiga, and Damico Anthony. Health coverage by race and ethnicity, 2010-2022, January 2024. URL `https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-coverage-by-race-and-ethnicity/`.

[41] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[42] Keiko Honda. Factors underlying variation in receipt of physician advice on diet and exercise: applications of the behavioral model of health care utilization. *American Journal of Health Promotion*, 18(5):370–377, 2004.

[43] Chanelle J Howe, Zinzi D Bailey, Julia R Raifman, and John W Jackson. Recommendations for using causal diagrams to study racial health disparities. *American Journal of Epidemiology*, 191(12):1981–1989, 2022.

[44] Lily Hu. What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, 21(1-2):1–26, 2023.

[45] John W Jackson. On the interpretation of path-specific effects in health disparities research. *Epidemiology*, 29(4):517–520, 2018.

[46] K Keisler-Stakey and Lisa N Bunch. Health insurance coverage in the United

States: 2020, 2021. URL https://www.census.gov/library/publications/2021/demo/p60-274.html.

[47] Raynard S Kington and James P Smith. Socioeconomic status and racial and ethnic differences in functional status associated with chronic diseases. *American Journal of Public Health*, 87(5):805–810, 1997.

[48] Naomi Y Ko, Susan Hong, Robert A Winn, and Gregory S Calip. Association of insurance status and racial disparities with the detection of early-stage breast cancer. *JAMA Oncology*, 6(3):385–392, 2020.

[49] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

[50] Theis Lange, Stijn Vansteelandt, and Maarten Bekaert. A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, 176(3):190–195, 2012.

[51] M Barton Laws, Yoojin Lee, William H Rogers, Mary Catherine Beach, Somnath Saha, P Todd Korthuis, Victoria Sharp, Jonathan Cohn, Richard Moore, and Ira B Wilson. Provider–patient communication about adherence to anti-retroviral regimens differs by patient race and ethnicity. *AIDS and Behavior*, 18:1279–1287, 2014.

[52] Benjamin Lê Cook, Thomas G McGuire, Kari Lock, and Alan M Zaslavsky. Comparing methods of racial and ethnic disparities measurement across different settings of mental health care. *Health Services Research*, 45(3):825–847, 2010.

[53] Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389 – 2415, 2019. doi: 10.1214/19-AOAS1282.

[54] Lei Liu, Ya-Chen Tina Shih, Robert L. Strawderman, Daowen Zhang, Bankole A. Johnson, and Haitao Chai. Statistical analysis of zero-inflated nonnegative continuous data. *Statistical Science*, 34(2), May 2019. ISSN 0883-4237. doi: 10.1214/18-STS681.

[55] Shiwani Mahajan, César Caraballo, Yuan Lu, Javier Valero-Elizondo, Daisy Massey, Amarnath R Annapureddy, Brita Roy, Carley Riley, Karthik Murugiah, Oyere Onuma, et al. Trends in differences in health status and health care access and affordability by race and ethnicity in the United States, 1999-2018. *Jama*, 326(7):637–648, 2021.

[56] Caleb H Miles. On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1154–1172, 2023.

[57] Caleb H Miles, Ilya Shpitser, Phyllis Kanki, Seema Meloni, and Eric J Tchetgen Tchetgen. On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107(1):159–172, 2020.

[58] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the Thirty Second Conference on Association for the Advancement of Artificial Intelligence (AAAI-32nd)*. AAAI Press, 2018.

[59] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR, 2019.

[60] National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The health consequences of smoking — 50 years of progress*. Reports of the Surgeon General. Centers for Disease Control and Prevention (US), Atlanta (GA), 2014.

[61] National Center for Health Statistics (US). *Volume of physician visits by place of visit and type of service*. Number 18. US Government Printing Office, 1965.

[62] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[63] Sungchul Park, Dylan H Roby, Jessie Kemmick Pintor, Jim P Stimpson, Jie Chen, and Alexander N Ortega. Insurance coverage and health care utilization among Asian youth before and after the Affordable Care Act. *Academic Pediatrics*, 20 (5):670–677, 2020.

[64] Judea Pearl. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–20, San Francisco, CA: Morgan Kaufmann, 2001.

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

[66] Judea Pearl. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, 6(2):20182001, 2018.

[67] Eric C Polley and Mark J Van der Laan. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2010.

[68] James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[69] Jack P Shonkoff, W Thomas Boyce, and Bruce S McEwen. Neuroscience, molecular biology, and the childhood roots of health disparities: building a new framework for health promotion and disease prevention. *Jama*, 301(21):2252–2259, 2009.

[70] Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)*, 37: 1011–1035, 2013.

[71] Ilya Shpitser and Eric Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of statistics*, 44(6):2433, 2016.

[72] Ilya Shpitser and Eric J. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016.

[73] Makiera Simmons, Kinfe G Bishu, Joni S Williams, Rebekah J Walker, Aprill Z Dawson, and Leonard E Egede. Racial and ethnic differences in out-of-pocket expenses among adults with diabetes. *Journal of the National Medical Association*, 111(1):28–36, 2019.

[74] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186(2):184–193, 2017.

[75] Mats J Stensrud, Jessica G Young, Vanessa Didelez, James M Robins, and Miguel A Hernán. Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 117(537):175–183, 2022.

[76] An-Shun Tai and Sheng-Hsuan Lin. Integrated multiple mediation analysis: A robustness-specificity trade-off in causal structure. *Statistics in Medicine*, 40(21): 4541–4567, 2021.

[77] An-Shun Tai, Le-Hsuan Liao, and Sheng-Hsuan Lin. On the conventional definition of path-specific effects: Fully mediated interaction with multiple ordered mediators. *Epidemiology*, 33(6):817–827, 2022.

[78] Eric J Tchetgen Tchetgen. Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine*, 32(26):4567–4580, 2013.

[79] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816, 2012.

[80] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[81] US Department of Health and Human Services. About the office of minority health, November 2019. URL `https://minorityhealth.hhs.gov/about-office-minority-health`.

[82] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

[83] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[84] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[85] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2014.

[86] Tyler J VanderWeele. Commentary: on causes, causal inference, and potential outcomes. *International Journal of Epidemiology*, 45(6):1809–1816, 2016.

[87] Tyler J. VanderWeele and Miguel A. Hernán. Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for nonmanipulable expo-

sures, with application to the effects of race and sex. In *Causality*, chapter 9, pages 101–113. John Wiley & Sons, Ltd, 2012. doi: 10.1002/9781119945710.ch9.

[88] Tyler J VanderWeele and Miguel A Hernan. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.

[89] Tyler J VanderWeele and Whitney R Robinson. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484, 2014.

[90] Stijn Vansteelandt, Maarten Bekaert, and Theis Lange. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1(1): 131–158, 2012.

[91] Jacob Wallace, Anthony Lollo, Kate A Duchowny, Matthew Lavallee, and Chima D Ndumele. Disparities in health care spending and utilization among Black and White Medicaid enrollees. *JAMA Health Forum*, 3(6):e221398–e221398, 2022.

[92] WHO Commission on Social Determinants of Health and World Health Organization. *Closing the gap in a generation: health equity through action on the social determinants of health: Commission on Social Determinants of Health final report.* World Health Organization, 2008.

[93] David R Williams, Naomi Priest, and Norman B Anderson. Understanding associations among race, socioeconomic status, and health: Patterns and prospects. *Health Psychology*, 35(4):407, 2016.

[94] Ziyue Wu, Seth A Berkowitz, Patrick J Heagerty, and David Benkeser. A two-stage super learner for healthcare expenditures. *Health Services and Outcomes Research Methodology*, 22(4):435–453, 2022.

[95] Jessica G Young, Lauren E Cain, James M Robins, Eilis J O'Reilly, and Miguel A Hernán. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences*, 3:119–143, 2011.

[96] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

[97] Xiang Zhou. Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):794–821, 2022.

[98] Xiang Zhou and Teppei Yamamoto. Tracing causal paths from experimental and observational data. *The Journal of Politics*, 85(1):250–265, 2023.

[99] Samuel H Zuvekas and Gregg S Taliaferro. Pathways to access: health insurance, the health care delivery system, and racial/ethnic disparities, 1996–1999. *Health Affairs*, 22(2):139–153, 2003.