UNIVERSITY OF CALIFORNIA

Los Angeles

Personalized and Situation-Specific Decision Making

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Scott Mueller

2025

ABSTRACT OF THE DISSERTATION

Personalized and Situation-Specific Decision Making

by

Scott Mueller
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2025
Professor Judea Pearl, Chair

Personalized and situation-specific decision making is a framework that optimizes both population-based and individual-based utilities by leveraging causal knowledge and counterfactual reasoning. This dissertation addresses the shortfalls of traditional decision strategies, which often overlook crucial individual heterogeneity. It develops novel methods of estimating how specific people or cases would respond under alternative actions or treatments, then uses those estimates to optimize decisions. By combining evidence from both experiments and observational studies, the approach side-steps the fundamental limitation of observing only one outcome per individual. As a result, an individual's probability of benefiting from, or being harmed by, an intervention can be estimated more precisely than previously possible.

The proposed methods also incorporate domain knowledge through causal models, and culminate in a practical way to handle non-binary ordinal outcomes, substantially expanding the usefulness of counterfactual reasoning. This framework is validated on real-world data from Tennessee's STAR project. Population-level data is distilled into individual-level probabilities, which in turn sharpen both broad policy decisions and personalized choices. I show

how the benefits of this framework extend across virtually every industry and discipline.

The dissertation of Scott Mueller is approved.

Eunice Jun

Chad Hazlett

Adnan Darwiche

Onyebuchi Arah

Judea Pearl, Committee Chair

University of California, Los Angeles

2025

To my advisor, Judea Pearl, who belongs among the greatest minds in history. Your relentless curiosity and energy continue to inspire me, and I am extraordinarily fortunate to know and learn from you. There is no upper bound on my respect and admiration for you, and I am forever indebted for the privilege of earning my PhD under your guidance.

To my kids, Ken and Margo, whom I was supposed to inspire. Turns out the causal direction was flipped.

To my wife, Antoaneta (Toni), who supported my decision to return to school mid-career and stuck it through. Your patience made this possible.

To my parents, thank you for supporting and helping the kids whenever you were able.

To my father-in-law, whose extraordinary support involved periodically traveling from Romania to stay with us, making my life easier and better.

To my doctoral committee, Professors Adnan Darwiche, Eunice Jun, Chad Hazlett, and Onyebuchi (Onyi) Arah, for their invaluable guidance. Adnan, serving as your TA four times was a privilege; your exceptional teaching deeply influenced my own approach to education. Eunice, your insights, feedback, and discussions have been incredibly helpful, and I look forward to future collaborations. Chad and Onyi, your outstanding causal inference courses were instrumental in kickstarting my education in causal inference and I'm so grateful to have been allowed to play a role in the Practical Causal Inference Lab.

To my academic brother, Ang Li, whose support started even before my first day at UCLA. Your groundbreaking unit selection work inspired me to explore Probabilities of Causation, profoundly influencing my research. I am grateful for your continued collaboration and feedback long after your graduation.

To my academic sister, Chi Zhang, who offered guidance at every stage of my UCLA journey, from feedback on papers and presentations to post-graduation career advice.

To Kaoru Mulvihill, for handling all administration, allowing me to focus on my work.

Lastly, to Rumen Iliev, whose leadership at the Toyota Research Institute (TRI) made our collaboration a pleasure. Your friendship is deeply valued.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Parts of Chapters 3 and 5 were adapted from [MP23c] and its addendum [MP23b], respectively. Professor Judea Pearl contributed to the writing of that paper and article. Chapter 4 is a version of [Mue21].

Six subsections within Sections 4.3 and 4.4 build upon and extend work done by Ang Li and myself. That work was written in [MLP22]. In particular, sections 4.3.2.2 and 4.3.3.2 advanced this work with conditions formalized under which the new P(benefit) bounds have superiority over the old PNS bounds, practical usage guidelines, and additional analyses, examples, and graphical criterion. Sections 4.4.1.1 and 4.4.2.1 similarly advanced this work with a further developed derivation on P(benefit) bounds with mediators, relaxed graphical criterion that is more flexible, and graphical methods to test applicability of formulas. Sections 4.3.4 and 4.3.5 are derived directly from our paper.

The first three sections of Chapter 7 are a version of my initial research in monotonicity, written in [MP23a]. Professor Judea Pearl contributed to the writing of that paper.

VITA

2012–2018 Founder and CEO, UCode, Hermosa Beach, California. UCode teaches kids and teenagers Computer Science. Over 10,000 students taught. $2.5 million raised from Bloomberg Beta, Idealab, DFJ Frontier, High Line Venture Partners, Fred Wilson, Brian Lee, and Klaus Schauser. Taught at 7 in-person centers and 20 schools in Southern California and New York. HTML/CSS/JavaScript, Elm, Computer Architecture, Swift, and Haskell curriculum.

2016–2019 Computer Science and Calculus Teacher, Ad Astra School @ SpaceX, Hawthorne, California. Created and taught rigorous Computer Science class, including Combinatory Logic and Lambda Calculus with "To Mock a Mockingbird" book and thorough introduction to the Haskell programming language, and rigorous proof-based Calculus class to high-achieving 11- to 14-year-old students. Taught Nand2Tetris parts I and II to 10- to 12-year-old students, and Scheme and UCode curriculum to all 7- to 14-year-old students.

2019–2020 Causal Inference Teacher, Ad Astra School @ SpaceX, Hawthorne, California. Created and taught rigorous Causal Inference class for 13- to 14-year-old students, including thorough introduction with "Causal Inference in Statistics: A Primer" book, along with instructional materials to scaffold book and make it accessible. Created and taught Psychology of Decision Making class for 12- to 14-year-old students, including biases and decision making with Kahneman's "Thinking, Fast and Slow" book, along with instructional materials to support book and rational thinking.

2021–present Graduate Student Researcher, Cognitive Systems Laboratory, Computer Science Department, UCLA.

2021–2021 Teaching Assistant, Computer Science Department, UCLA. Graduate course CS 262A: Learning and Reasoning with Bayesian Networks

2022–present Web Developer and Co-Manager, Practical Causal Inference Lab, UCLA.

2022–2024 Teaching Assistant, Computer Science Department, UCLA. Graduate course CS 264A: Automated Reasoning: Theory and Applications

2024–present Software Architect, Sup AI, Hermosa Beach, California. Responsible for all software development and cloud infrastructure. Deployed robust AI infrastructure to power chat and content generation. Implemented ensemble method with multiple LLMs, search, and RAG.

PUBLICATIONS

[MP19] Scott Mueller and Judea Pearl. "Fréchet Inequalities – Visualization, Applications, and History," Causal Analysis in Theory and Practice, November 2019. https://causality.cs.ucla.edu/blog/index.php/2019/11/05/frechet-inequalities/.

[MP20] Scott Mueller and Judea Pearl. "Which Patients Are in Greater Need: A Counterfactual Analysis with Reflections on COVID-19," Causal Analysis in Theory and Practice, April 2020. https://causality.cs.ucla.edu/blog/index.php/2020/04/02/which-patients-are-in-greater-need-a-counterfactual-analysis-with-reflections-on-covid-19/.

[MP21] Scott Mueller and Judea Pearl. "Personalized Decision Making," Causal Analysis in Theory and Practice, April 2021. https://causality.cs.ucla.edu/blog/index.php/2021/04/29/personalized-decision-making/.

[MLP22] Scott Mueller, Ang Li, and Judea Pearl. "Causes of Effects: Learning Individual Responses from Population Data," In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pp. 2712–18. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, July 2022. https://doi.org/10.24963/ijcai.2022/376.

[FM22] Andrew Forney and Scott Mueller. "Causal Inference in AI Education: A Primer," Journal of Causal Inference, vol. 10, no. 1, pp. 141–73, July 2022. https://doi.org/10.1515/jci-2021-0048.

[LMP23] Ang Li, Scott Mueller, and Judea Pearl. "ϵ-Identifiability of Causal Quantities," January 2023.

[MP23c] Scott Mueller and Judea Pearl. "Personalized Decision Making – A Conceptual Introduction," Journal of Causal Inference, vol. 11, no. 1, January 2023. https://doi.org/10.1515/jci-2022-0050.

[MP23b] Scott Mueller and Judea Pearl. "Personalized Decision Making under Concurrent-Controlled RCT Data," Causal Analysis in Theory and Practice, March 2023. https://causality.cs.ucla.edu/blog/index.php/2023/03/17/personalized-decision-making-under-concurrent-controlled-rct-data/.

[MP23a] Scott Mueller and Judea Pearl. "Monotonicity: Detection, Refutation, and Rami-

fication," 2023 RAND Center for Causal Inference (CCI) Symposium, August 2023.

[ZLM24] Chi Zhang, Ang Li, Scott Mueller, and Rumen Iliev. "Causal AI Framework for Unit Selection in Optimizing Electric Vehicle Procurement," AAAI 2024 Workshop on Sustainable AI, Vancouver, Canada, February 2024.

[MP24] Scott Mueller and Judea Pearl. "The Meaning of 'Harm' in Personalized Medicine – An Alternative Perspective," American Journal of Epidemiology, November 2024. https://doi.org/10.1093/aje/kwae441.

GLOSSARY

ATE Average Treatment Effect; average effect of a treatment across a population.

CATE Conditional Average Treatment Effect; ATE conditioned on one or more covariates.

CoE Causes of Effects; counterfactual probabilities of causation.

DAG Directed Acyclic Graph; graph whose nodes are connected by directed edges and where its edges do not form any directed cycles or loops.

EoC Effects of Causes; probability or expectation of an outcome given an interventional or observational cause.

Exogeneity X is exogenous for Y iff the way Y would potentially respond to experimental conditions x or x' is independent of the actual value of X.

ITE Individual Treatment Effect; difference in outcomes that an individual would attain under different decisions.

LoTP Law of Total Probability; $\sum_{\mathbf{b}\in\mathbf{B}} P(\mathbf{a}, \mathbf{b}) = P(\mathbf{a})$.

MITE Monotonic Incremental Treatment Effect; stronger version of monotonicity, where treatment X does not decrease the level of Y and increases Y by at most one level.

PN Probability of Necessity; $P(y'_{x'}|x, y)$.

PNS Probability of Necessity and Sufficiency; $P(y_x, y'_{x'})$, also known as Probability of Benefit (P(benefit)).

PoC Probability of Causation; counterfactual probability of an event being the cause of an effect.

PoCs Probabilities of Causation; plural of Probability of Causation.

PS Probability of Sufficiency; $P(y_x|x', y')$.

RCT Randomized Controlled Trial; experimental research study design where participants are randomly assigned to different groups (treatment or control) to assess the effectiveness of an intervention or treatment.

Utility Usefulness, value, preference level, or profit that is placed on something.

CHAPTER 1

Introduction

1.1 Background

Decision-making systems across domains like healthcare, business, and policy typically rely on averages and broad trends. However, what works *on average* for a group may not be optimal for a specific individual or situation. Moreover, average effects rarely incorporate the utility associated with different counterfactual outcomes, frequently turning seemingly optimal decisions into severely suboptimal ones. This utility utilization frequently turns what appears to be an optimal decision into a severely suboptimal one. Personalized and situation-specific decision-making focuses on tailoring choices to the unique characteristics of a person or unit *and* the context at hand. The motivation for this approach stems from everyday observations: one patient might respond very well to a treatment that yields only modest benefits in the average clinical trial, or a marketing incentive that usually boosts sales might backfire for certain customers. Traditional population-based decision policies, which optimize the outcome for an average subgroup, can obscure these nuances.

Consider a motivating example from public health. Imagine a scenario in which a new vaccine is in short supply during a pandemic, and officials must decide who should be vaccinated first. An initial causal analysis might find that elderly patients have a higher average survival rate with the vaccine versus without the vaccine (70%) compared to younger patients (60%), suggesting that the elderly benefit more. This is a classic population-level conclusion. Yet, a closer, individualized look reveals a different story. We want to know who would

survive if vaccinated but *not* survive if *not* vaccinated. Answering this question requires counterfactual reasoning: comparing an outcome for the same person under two alternative choices (vaccinate vs. not vaccinate). Since we cannot directly observe both outcomes for any single individual, we must infer them through data and models. When researchers applied such counterfactual analysis to the vaccine scenario, they found a puzzling result. It was not at all clear which age group should be prioritized. The probability that an elderly person's life is saved *because* of the vaccine (and would be lost without it) had such wide uncertainty (70% to 85%) that it overlapped with that of a younger person (60% to 80%). In fact, with additional real-world data, the analysis indicated that the younger group, in reality, had the higher benefit probability (75% to 80% versus 70% to 74%), flipping the policy recommendation. This example highlights the crux of the problem. A decision policy based purely on group averages can be misleading, whereas a personalized, situation-aware analysis can unveil insights that change who we decide to help first.

The above scenario underscores why this research is necessary. Many high-stakes decisions in medicine, economics, and public policy suffer from the fundamental limitation that we only see one outcome per individual. We rarely get to rewind the clock and see counterfactual outcomes (e.g., how a patient would have fared under an alternative treatment). As a result, conventional analytics give us population effects, such as an Average Treatment Effect (ATE) from a trial, but cannot directly tell us what would happen to a specific person. Counterfactual reasoning provides the language and tools to bridge this gap by estimating quantities like an individual's probability of benefit or harm from an action. However, early methods to do this were often impractical. Without strong assumptions or knowledge of the underlying data generating process, one could only derive very wide bounds on an individual's benefit probability, too broad to guide decisions. This dissertation builds a foundation to overcome those hurdles. Leveraging recent advances in causal inference, I integrate multiple data sources and domain knowledge to estimate individual-specific effects more precisely than previously possible. By combining experimental trial data with obser-

vational data, valuable information about individual behavior is obtained that a randomized trial alone would miss. Furthermore, by incorporating causal models of the decision scenario, structural knowledge is added that helps refine estimates even more. In sum, I use *reasonable assumptions*, including all available evidence and causal structure, to answer "What is the likelihood that an individual responds in a specific way if we do X and, *at the same time*, this individual responds in a specific alternative way if we do an alternative to X?" as reliably as possible.

Ultimately, the significance of personalized and situation-specific decision-making is its potential to improve outcomes across a wide range of fields. I will show how better estimates of counterfactual Probabilities of Causation (PoCs) in a heterogeneous population leads to decisions that are not just more optimal for a particular set of preferences, but are also fairer, more efficient, and more transparent.

1.2 Decisions

Decision making underpins progress in domains ranging from education, healthcare, and public policy to choosing what to eat for dinner. In almost every facet of life, we are confronted with the task of choosing an option that leads to the best future outcome.

Early approaches relied on predictive models, using historical data to forecast outcomes. This is effective for trends but blind to causality. Unfortunately, predictive decision making is still prevalent in industry, academia, and certainly among people not trained in causal inference. This makes sense as our intuition tells us to predict the future for each option and select the one that seems to lead to the best future. Conditional probabilities tell us how often something happens, but does not tell us what will happen if we intervene. The supplement industry, valued at over half a trillion dollars in 2025 [Res], thrives on decisions based on observations like, "People taking this supplement seem super healthy, so I'll buy it." Of course, many of those supplement buyers are doing plenty of other things to make

themselves healthy, such as exercising. This is level 1 thinking in the Pearl Causal Hierarchy [BCI22]. Decision making is fundamentally causal. We should not be making predictions, we should be understanding the consequences of actions.

The advent of interventional methods, such as Randomized Controlled Trials (RCTs), shifted focus to causal effects, typically measured via the ATE. This was a big step forward. A decision is an intervention. Choices represent potential actions. To make good decisions, we need to understand what would happen if we intervene. However, while robust for population-level insights, ATE masks individual variability. This is a critical shortfall when decisions must suit specific people, situations, or contexts [MP23c]. Interventional methods correspond to level 2 in the Pearl Causal Hierarchy.

Modern decision challenges demand counterfactual reasoning. This dissertation advances this paradigm through PoCs, measures of an action's causal impact on individuals (e.g., benefit or harm), building upon frameworks like those in Tian and Pearl's work [TP00]. Unlike ATE, PoCs capture individual heterogeneity, but their estimation is complicated by the fundamental problem of causal inference: only one outcome per individual is ever observable.

Optimal decision making involves three key components:

1. Determine utility of actions on each counterfactual response type
2. Estimate counterfactual probabilities
3. Apply utilities to counterfactual probabilities
4. Choose action with highest utility

where utility is the usefulness, value, preference level, or profit that is placed on something. A counterfactual response type is how a unit simultaneously responds to different actions. For example, a patient would recover with medicine A and, simultaneously, remain sick with

medicine B. I will tackle the first three optimal decision-making components in the rest of this dissertation, leaving the reader with the challenge of the fourth component.

1.3 Contributions

The work presented in this dissertation makes several novel contributions to the theory and practice of decision science.

In Chapter 3, the counterfactual probabilities focused on in this dissertation are introduced. These PoCs are analyzed mathematically and intuitively, providing insights into how they can be estimated and used. Chapter 9 will expand on these PoCs with non-binary outcomes, greatly enhancing their usefulness.

Chapters 4 and 5 present methods to estimate PoCs. Because to their counterfactual nature, PoCs have historically been difficult to estimate. However, with reasonable assumptions, domain knowledge, causal structure, or covariate date, estimates can be precise enough to make better decisions from.

Chapter 6 presents another way, potentially in combination with other techniques in this dissertation and elsewhere, to narrow bounds and better estimate PoCs. The insight here is to use the intention to treat or not treat as evidence. Once a decision has been made, but before it is actually enforced, probabilities change, which could affect decisions before it is too late.

Monotonicity is an assumption that outputs cannot decrease with increasing inputs. If this assumption can legitimately be made, many causal inference problems become simple. Chapter 7 presents tests for necessary conditions on data in order for monotonocity to possibly hold and sufficient conditions on data for monotonicity to definitely hold. Consequences of violations of monotonicity are explored. This leads to potential better estimates on PoCs.

Similar to monotonicity, selection bias can be difficult to detect or refute. Unfortunately,

selection bias can also be difficult to avoid. Chapter 8 quantifies severity of selection bias, derives its bounds, and incorporates those bounds in causal analysis.

Finally, a cornerstone contribution is extending PoC to non-binary ordinal outcomes, moving beyond binary success and failure. In Chapter 9, new bounds and estimates, along with conditions for point estimates, are derived. Existing concepts, such as ATE and monotonicity, are analyzed and adjusted in light of non-binary outcomes. While non-binary outcomes increase decision-making complexity, several techniques help manage it. The final chapter and the dissertation end with a simple benefit function crucial for making optimal decisions using counterfactual reasoning with non-binary ordinal outcomes. That formula can be point estimated if the utility matrix passes a basic mathematical test. An algorithm is presented to then compute the overall utility of a decision using counterfactual reasoning.

A real-world dataset will be used as a practical illustration where appropriate. This includes chapter 9, where the outcome is discretized into binary and quaternary levels and compared. This will demonstrate the value of these new decision-making methods with counterfactual reasoning.

CHAPTER 2

Real-World Dataset (STAR)

2.1 Introduction

This dissertation leverages data from the STAR-and-Beyond database [ABB08], a longitudinal educational experiment initiated by the Tennessee State Department of Education in 1985. The original experiment tracked a cohort of students from kindergarten through third grade (1985–1989), collecting annual data on student achievement and related measures. Although the study continued gathering data into high school, this dissertation focuses solely on the primary cohort (grades K–3).

The experimental phase included 11,601 students who participated for at least one year. The dataset captures:

- Demographic characteristics
- School and classroom identifiers
- School and teacher attributes
- Experimental conditions ("class types")
- Norm- and criterion-referenced achievement test scores
- Motivation and self-concept assessments

To enhance causal analyses by combining experimental and observational data, records from 1,780 students in 21 non-STAR comparison schools (matched demographically to STAR

schools) are utilized. These observational data, however, are limited by fewer demographic and experimental condition details.

2.2 Overview and Goals of Project STAR

In May 1985, the Tennessee legislature initiated Project STAR (House Bill 544) to study the impact of reduced class sizes on primary school student achievement. The legislature mandated three key research questions:

- What effect does smaller class size have on student achievement and development in grades K–3?
- Do effects accumulate over multiple years in small classes compared to a single year?
- Does specialized teacher training for smaller classes or teacher aides influence student outcomes compared to teachers without such training

Small classes were defined as having 13–17 students. The study involved 79 schools across 42 districts, encompassing a diverse mix of inner-city, suburban, urban, and rural schools.

2.3 Selection Bias Considerations

Due to deliberate sampling methods, STAR schools were slightly larger and had marginally lower initial math and reading scores than statewide averages. Figure 2.1 compares the distribution of initial math scaled scores in STAR and non-STAR schools, illustrating this mild selection bias.

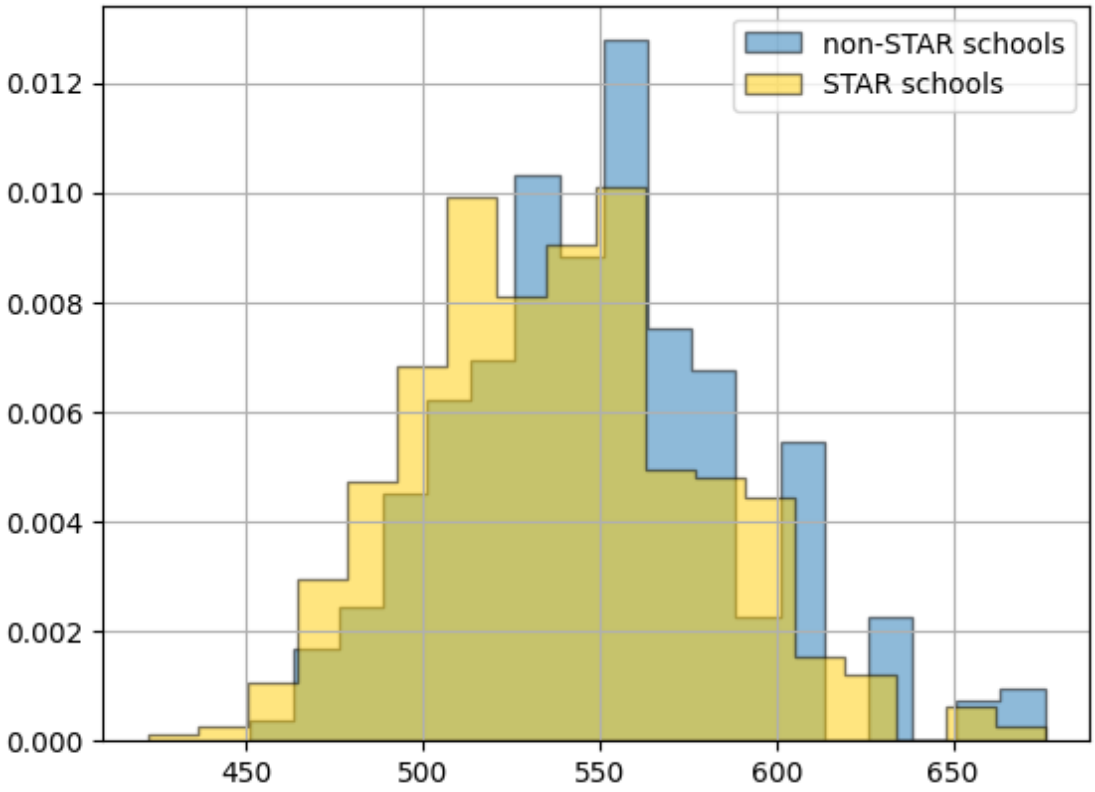


Figure 2.1: Total math scaled scores among STAR schools and non-STAR schools.

2.4 Observational Comparison Schools

Twenty-one non-participating schools, matched to STAR schools across 13 STAR school districts, formed a comparison group. These schools did not implement class-size interventions but administered identical achievement tests in grades 1–3 during the study period. Initial comparisons on academic achievement measured in second grade, prior to the experiment's start, confirmed that STAR and comparison schools were closely aligned in performance [WJB90, Table I-4]. However, unlike the randomized classroom assignments used in STAR schools, the comparison schools employed traditional, often non-random methods. Differences between these assignment strategies were analyzed by Zaharias, Achilles, and Cain [ZAC95].

2.5 Experimental Design

The STAR experiment followed one student cohort over four years, beginning with randomly assigning students entering kindergarten (1985) or first grade (1986) into three conditions:

1. Small class (13–17 students)
2. Regular class (22–25 students)
3. Regular class with a full-time teacher aide (22–25 students)

Initially, 128 small classes, 101 regular classes, and 99 regular-aide classes were formed. Kindergarten was not compulsory at the time, so many students entered in first grade. Students remained in their assigned condition throughout the project. Overall, 26.6% participated all four years, while another 22.0% joined in first grade and stayed through grade three.

2.6 Binary Treatment Simplification

There were initially three treatment levels, as described above: small class, regular class, and regular class with a teacher aide. After the kindergarten year showed negligible differences between regular classes and aide-assisted regular classes, these two groups were combined randomly into a single "regular" class type. About half of the students initially in regular classes were randomly moved to teacher-aide classes for subsequent years, and vice versa. No students were intentionally moved into or out of small classes.

2.7 Data

This dissertation employs the primary STAR cohort as experimental data and matched non-project schools as observational data. Unfortunately, the observational dataset contained far

fewer covariates. Therefore, the analyses in this dissertation focuses on second-grade total math scaled scores as the outcome variable M, class size as the treatment S, and first-grade total math scaled scores as the sole confounding variable F. This is diagrammed in Figure 2.2.

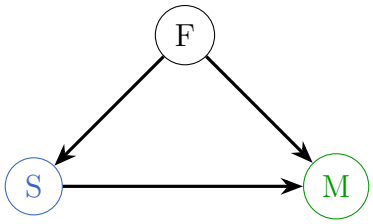


Figure 2.2: DAG for class size (S), first grade math score (F), and second grade math score (M).

Since the STAR consortium found no achievement differences between regular and teacher-aide classes, as described above, these two class types were randomly combined. Therefore, ternary treatment became binary.

Total math scale scores from the Stanford Achievement Test (SAT) ranged from 422 to 676. Those were discretized into binary and quaternary values. The specific cut points are shown in Table 2.1. First grade math scores were only transformed with quaternary discretization, while second grade math scores were transformed with both binary and quaternary discretizations. This allowed comparing binary results with the non-binary ordinal outcome methods presented in Chapter 9.

Range	Quaternary Discretization	Binary Discretization
422 to 487	0	0
488 to 550	1	
551 to 613	2	1
614 to 676	3	

Table 2.1: SAT math scale score discretization

One simplifying assumption was made due to the very small number of students in *small* classes. After removing problematic rows (primarily missing fields for second grade class size, first grade math score, and second grade math score), there were only 15 students in the comparison schools (observational data) in small classrooms during second grade. The assumption is, on average, that all students can be combined, from both observational and experimental datasets, to create the conditional probability table (CPT) of first grade math scores. This eliminated the problem of only 15 students in small classrooms. Although this assumption is significant, the observational data closely aligns with experimental data for this CPT, suggesting the impact on the analysis would be minimal. The DAG in Figure 2.2 also aligns with this strategy. The resulting distribution of grade 1 math scores is shown in Table 2.2 and Figure 2.3.

Grade 1 Math Score (F)	$P(F)$
0	0.1011
1	0.5303
2	0.3375
3	0.0311

Table 2.2: CPT for F: grade 1 math scores.

As a comparison, Table 2.3 and Figure 2.4 show the distribution of math scores among second graders. The distributions between first and second graders differ notably, with second graders performing considerably better on average. This increase was seen in both the limited data among the non-STAR project comparison schools and the STAR schools. Thus, the experimental intervention itself appeared not to substantially affect student scores. Participation in the study, even as a comparison school, may have been a factor.

Due to the small number of students in small class sizes in the comparison non-STAR schools, there were no students in small classes that had either the lowest or the highest quarter (according to the discretization in Table 2.1) of math scores in first grade. Table 2.4

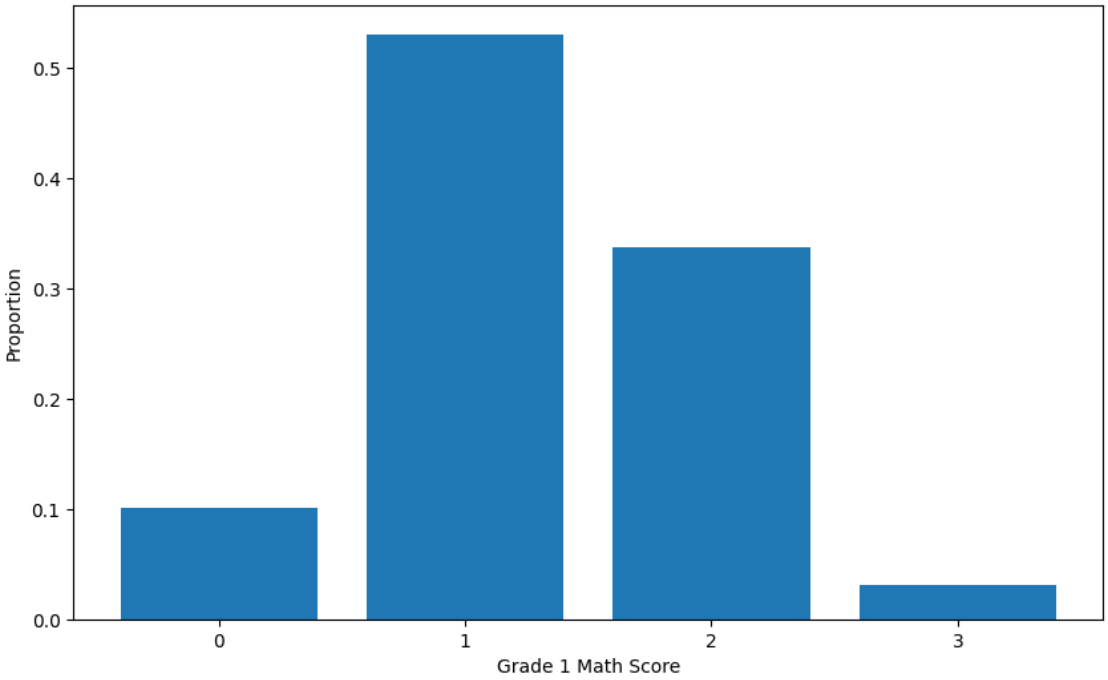


Figure 2.3: Histogram of discretized grade 1 math scores.

and Figure 2.5 show this.

To overcome this issue, an additional assumption was made. For non-STAR schools, class sizes were given for each class instead of class type (small or regular). Originally, I took the same ranges of students as the STAR schools to categorize students into small (13 to 17 students) or regular (22 to 25 students) size classes. Increasing the upper limit for small class size from 17 to 20 produced a more reasonable distribution of students. Like the assumption above about combining observational and experimental first grade math scores, this had negligible effect on the small number of students with grade 1 math scores of 1 and 2. Therefore, this appears to be a reasonable assumption. The updated CPT and chart are in Table 2.5 and Figure 2.6.

For the STAR schools, the distribution of second grade class sizes did not depend on first grade math scores at all. This is because the STAR school data is experimental. However, randomization did not produce evenly sized groups. Table 2.6 and Figure 2.7 show the

Grade 2 Math Score (M)	$P(M)$
0	0.0083
1	0.2155
2	0.5209
3	0.2330

Table 2.3: Distribution of grade 2 math scores.

Grade 1 Math (F)	$P(S = \text{regular})$	$P(S = \text{small})$
0	1.0000	0.0000
1	0.9594	0.0406
2	0.9622	0.0378
3	1.0000	0.0000

Table 2.4: Original CPT for S: proportion of students per grade 2 class size in non-STAR schools.

distribution.

This dissertation will require causal effect probabilities of the form $P\left(Y = y|do(X = x)\right)$. Tables 2.7 and 2.8 and Figures 2.8 and 2.9 show the distribution of STAR school second grade math scores (quaternary and binary discretized) by which class size they were randomly assigned to. This corresponds to the probability $P\left(M = m|do(S = s)\right)$.

Tables 2.9, 2.10, 2.11, and 2.12 and Figures 2.10, 2.11, 2.12, and 2.13 represent the quaternary and binary discretized second grade math scores for STAR schools (experimental data) and non-STAR schools (observational data).

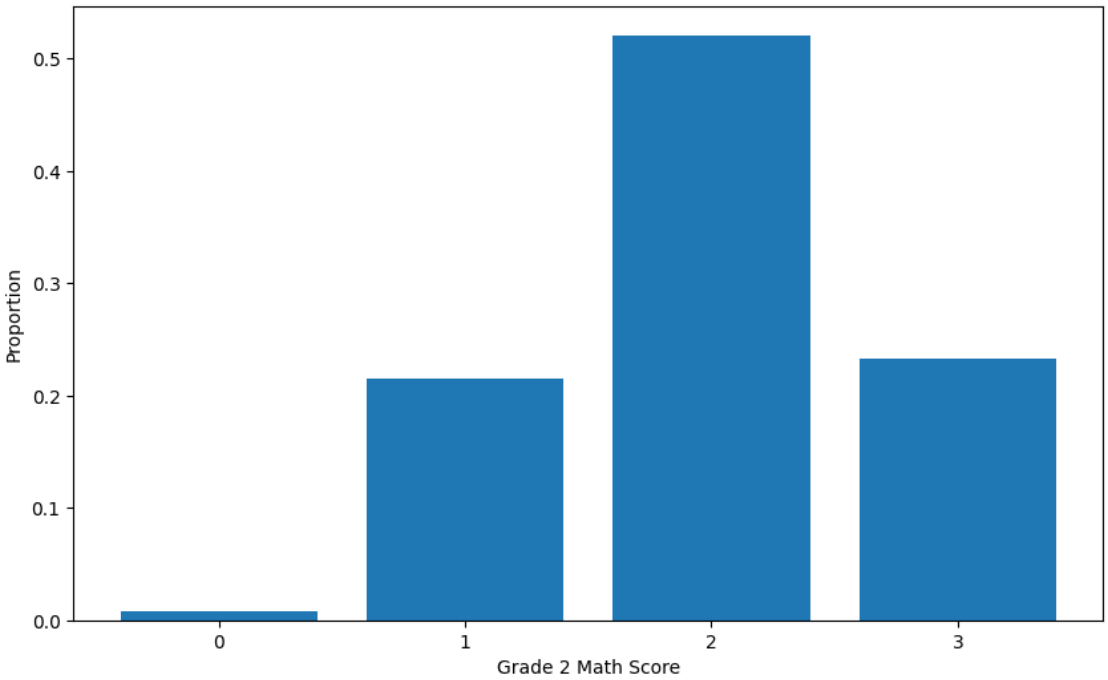


Figure 2.4: Histogram of discretized grade 2 math scores.

Grade 1 Math (F)	$P(S = \text{regular})$	$P(S = \text{small})$
0	0.8846	0.1154
1	0.7875	0.2125
2	0.8128	0.1872
3	0.8750	0.1250

Table 2.5: CPT for S: proportion of students per grade 2 class size in non-STAR schools after increasing upper limit of class size for small classes from 17 to 20.

Grade 2 Class Size (S)	$P(S)$
regular	0.6778
small	0.3222

Table 2.6: Experimental distribution for S: proportion of students per grade 2 class size in STAR schools.

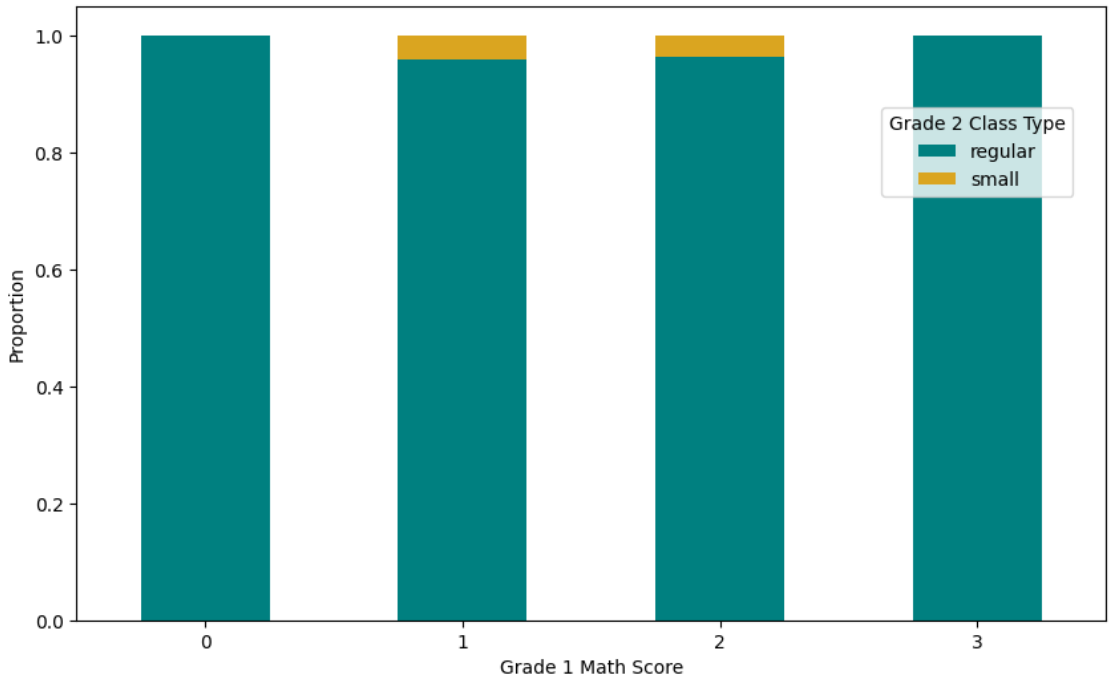


Figure 2.5: Stacked histogram of proportion of students per original grade 2 class size in non-STAR schools.

Grade 2 Class Type (S)	$P(0)$	$P(1)$	$P(2)$	$P(3)$
regular	0.0103	0.2437	0.5306	0.2154
small	0.0062	0.2023	0.5290	0.2624

Table 2.7: Experimental CPT for M: proportion of students with particular grade 2 math scores (quaternary discretized) per grade 2 class size in STAR schools.

Grade 2 Class Type (S)	$P(0)$	$P(1)$
regular	0.2540	0.7460
small	0.2086	0.7914

Table 2.8: Experimental CPT for M: proportion of students with particular grade 2 math scores (binary discretized) per grade 2 class size in STAR schools.

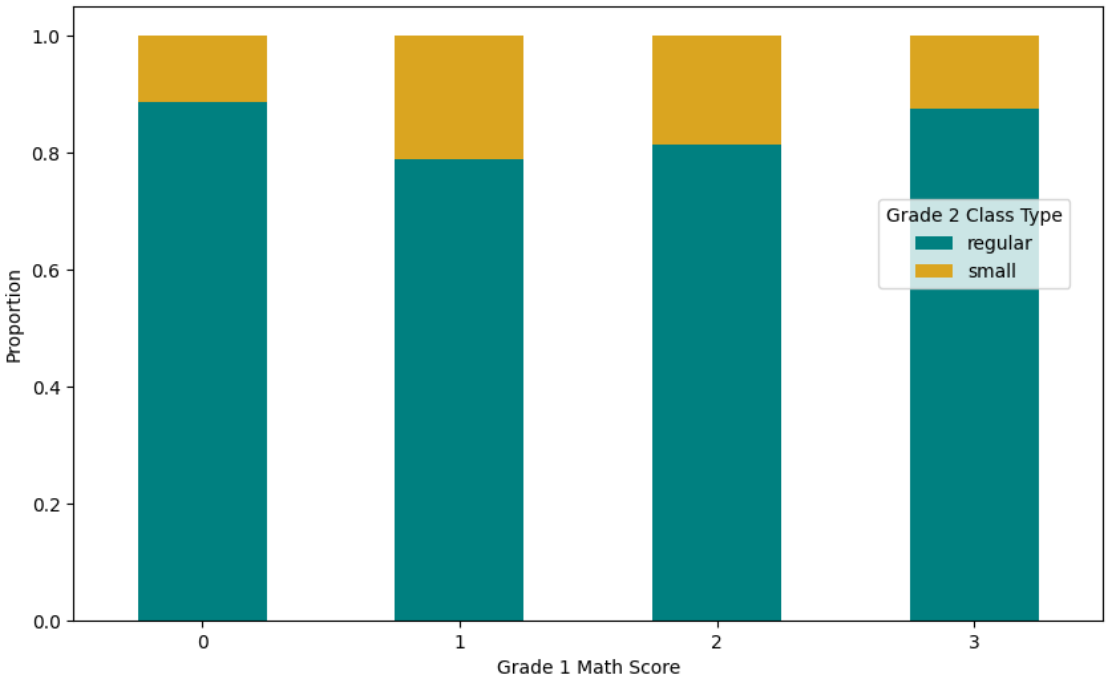


Figure 2.6: Stacked histogram of proportion of students per grade 2 class size in non-STAR schools.

Grade 1 Math (F)	Grade 2 Class (S)	$P(0)$	$P(1)$	$P(2)$	$P(3)$
0	regular	0.0714	0.6905	0.2302	0.0079
	small	0.0531	0.6726	0.2743	0.0000
1	regular	0.0029	0.2793	0.6227	0.0950
	small	0.0039	0.2584	0.6382	0.0995
2	regular	0.0000	0.0178	0.5058	0.4764
	small	0.0000	0.0333	0.4549	0.5118
3	regular	0.0000	0.0000	0.1833	0.8167
	small	0.0000	0.0000	0.1765	0.8235

Table 2.9: Experimental CPT for M: proportion of students with particular grade 2 math scores (quaternary discretized) per grade 1 math score and grade 2 class size in STAR schools.

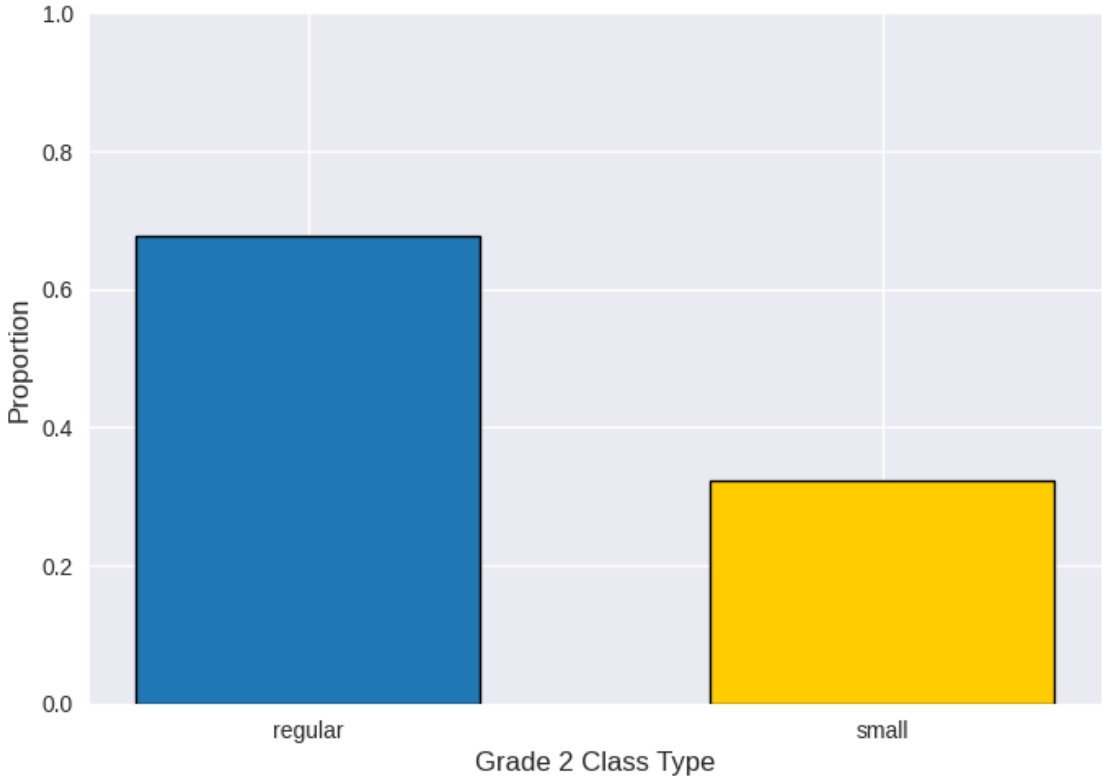


Figure 2.7: Histogram of students in grade 2 class sizes in STAR schools.

Grade 1 Math (F)	Grade 2 Class (S)	$P(0)$	$P(1)$
0	regular	0.7619	0.2381
	small	0.7257	0.2743
1	regular	0.2822	0.7178
	small	0.2623	0.7377
2	regular	0.0178	0.9822
	small	0.0333	0.9667
3	regular	0.0000	1.0000
	small	0.0000	1.0000

Table 2.10: Experimental CPT for M: proportion of students with particular grade 2 math scores (binary discretized) per grade 1 math score and grade 2 class size in STAR schools.

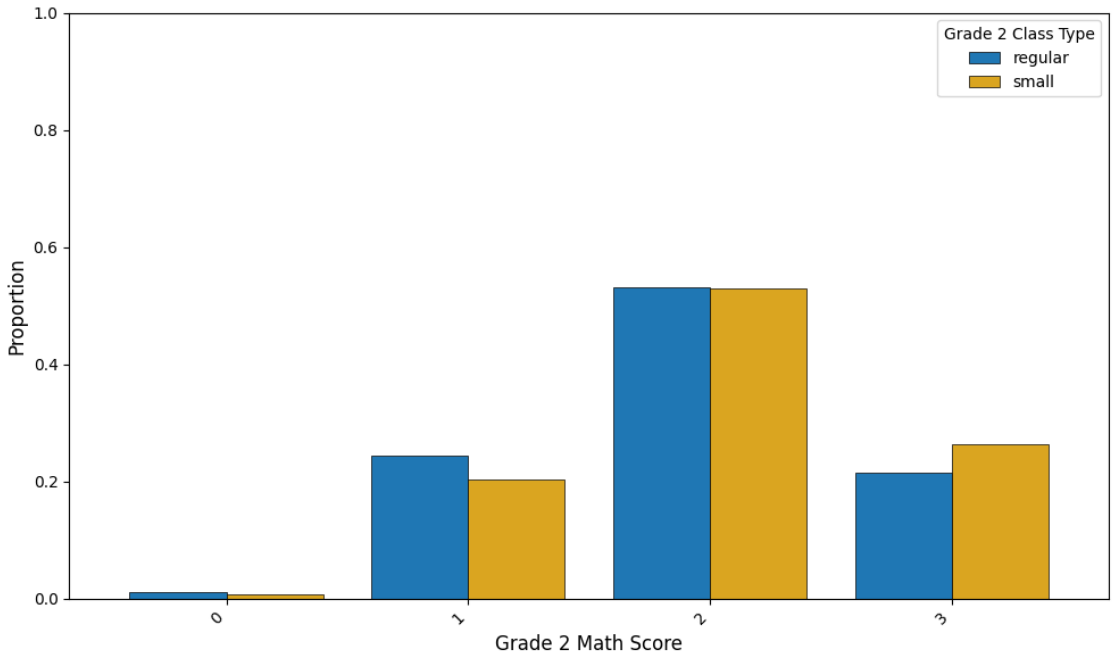


Figure 2.8: Side-by-side histogram of grade 2 math scores (quaternary discretized) by grade 1 score and class type in STAR schools.

Grade 1 Math (F)	Grade 2 Class (S)	$P(0)$	$P(1)$	$P(2)$	$P(3)$
0	regular	0.0000	0.4091	0.5909	0.0000
	small	0.0000	0.6667	0.3333	0.0000
1	regular	0.0054	0.1075	0.6774	0.2097
	small	0.0000	0.3137	0.5686	0.1176
2	regular	0.0000	0.0636	0.4624	0.4740
	small	0.0000	0.0789	0.5526	0.3684
3	regular	0.0000	0.0556	0.2222	0.7222
	small	0.0000	0.0000	0.0000	1.0000

Table 2.11: Observational CPT for M: proportion of students with particular grade 2 math scores (quaternary discretized) per grade 1 math score and grade 2 class size in non-STAR schools.

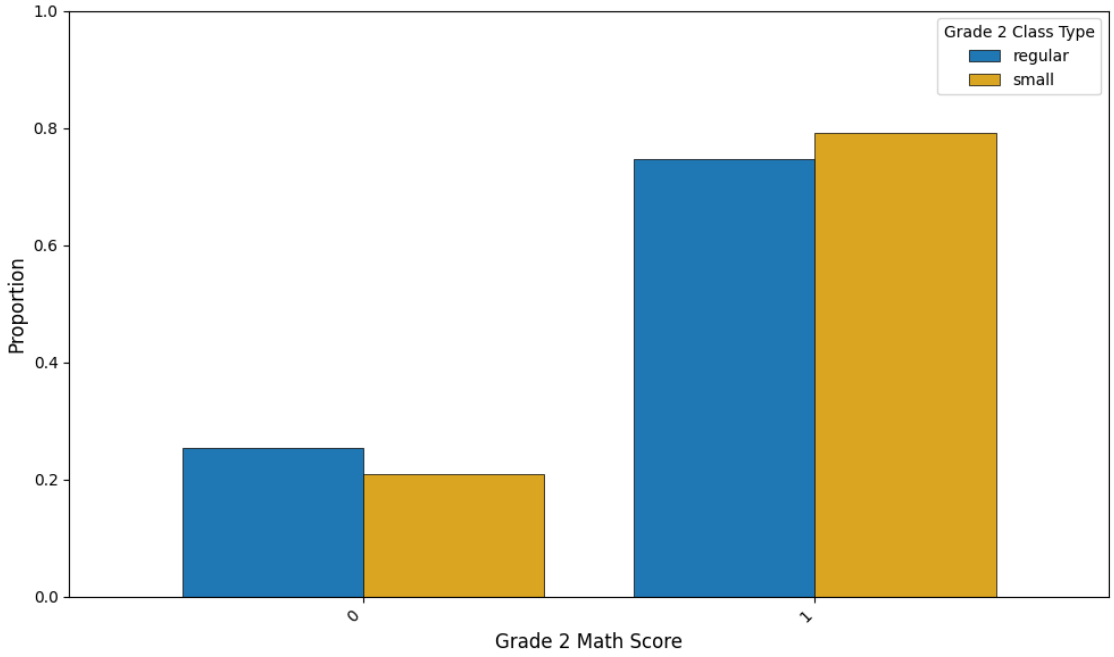


Figure 2.9: Side-by-side histogram of grade 2 math scores (binary discretized) by grade 1 score and class type in STAR schools.

Grade 1 Math (F)	Grade 2 Class (S)	$P(0)$	$P(1)$
0	regular	0.4091	0.5909
	small	0.6667	0.3333
1	regular	0.1129	0.8871
	small	0.3137	0.6863
2	regular	0.0636	0.9364
	small	0.0789	0.9211
3	regular	0.0556	0.9444
	small	0.0000	1.0000

Table 2.12: Observational CPT for M: proportion of students with particular grade 2 math scores (binary discretized) per grade 1 math score and grade 2 class size in non-STAR schools.

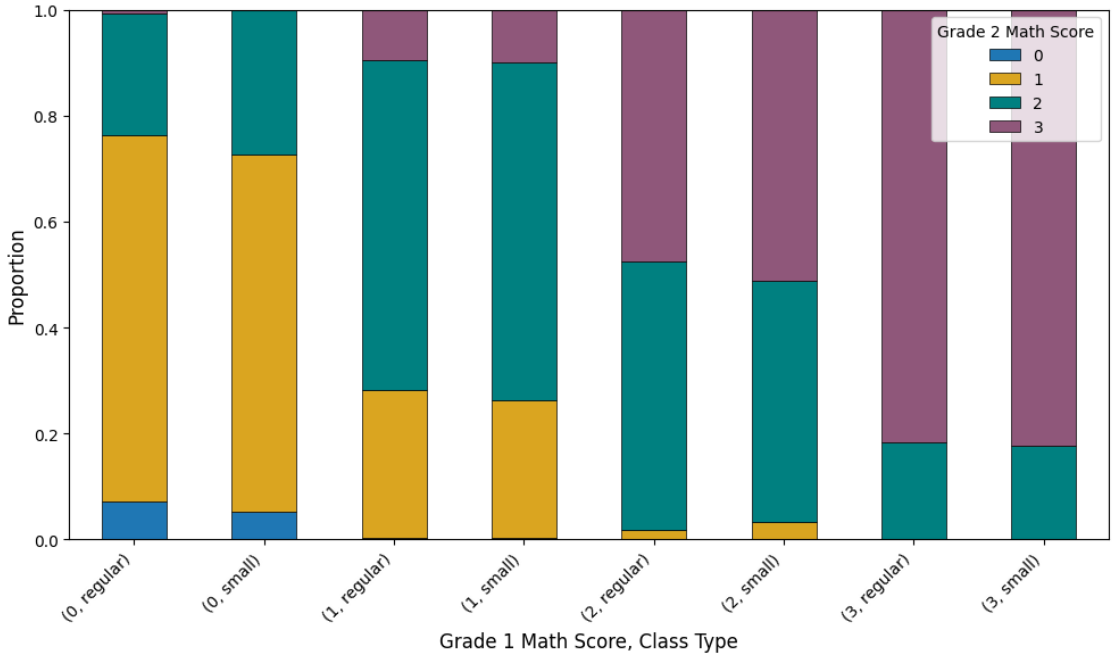


Figure 2.10: Stacked histogram of grade 2 math scores (quaternary discretized) by grade 1 score and class type in STAR schools.

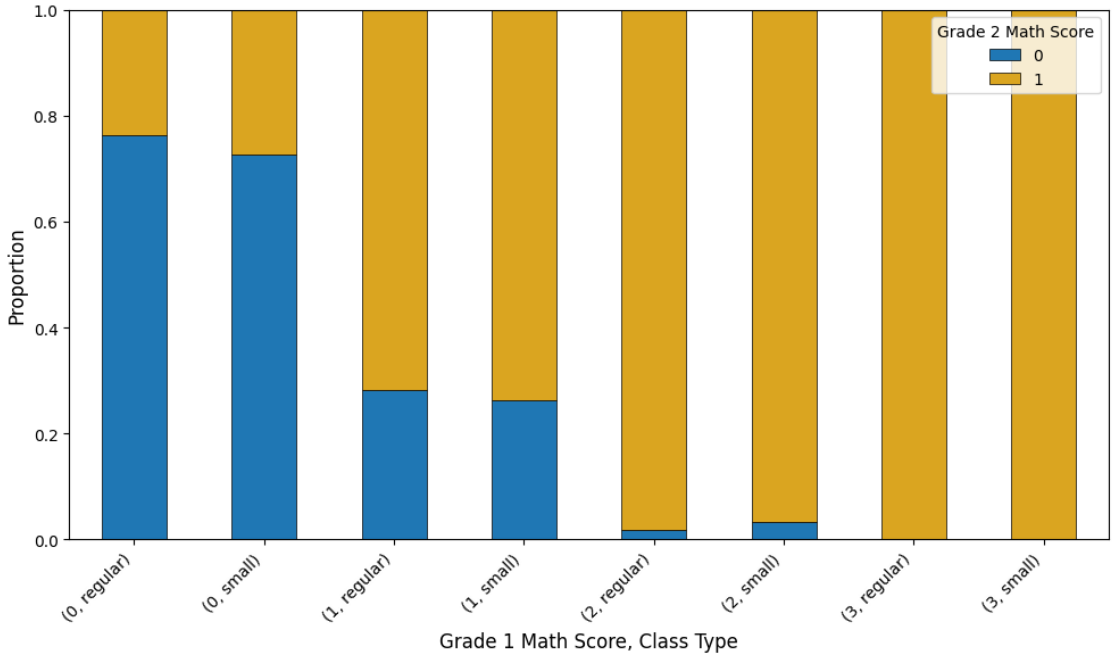


Figure 2.11: Stacked histogram of grade 2 math scores (binary discretized) by grade 1 score and class type in STAR schools.

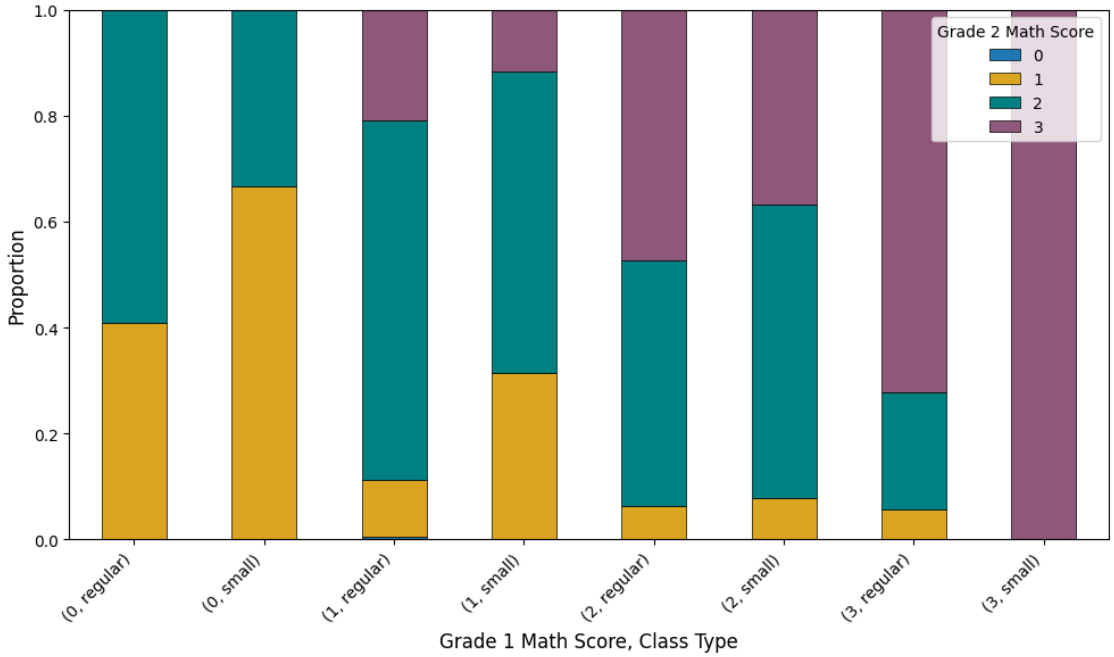


Figure 2.12: Stacked histogram of grade 2 math scores (quaternary discretized) by grade 1 score and class type in non-STAR schools.

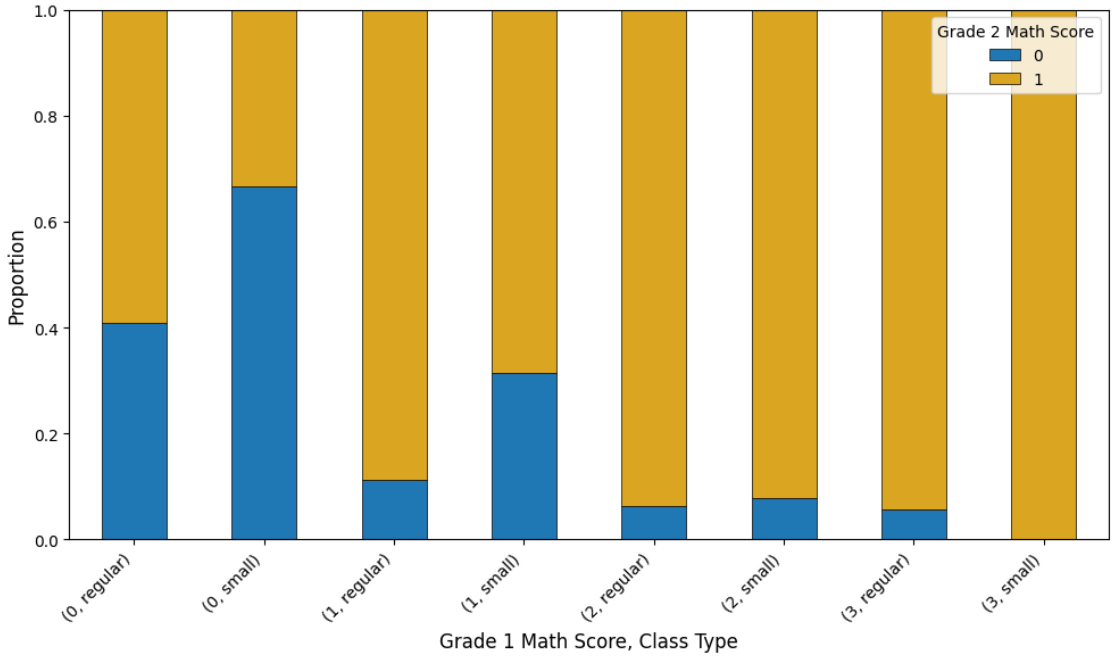


Figure 2.13: Stacked histogram of grade 2 math scores (binary discretized) by grade 1 score and class type in non-STAR schools.

CHAPTER 3

Probabilities of Causation

3.1 Introduction

This chapter examines the distinction between personalized and population-based decision-making and demonstrates the advantages of the former and how it can be achieved.

Formally, personalized decision making optimizes the Individual Treatment Effect (ITE) in Definition 3.1.1.

Definition 3.1.1 (Individual Treatment Effect (ITE))**.**

$$ITE(u) \triangleq Y(1,u) - Y(0,u) \tag{3.1}$$

where $Y(x,u)$ stands for the outcome that individual u would attain had decision $x \in \{1,0\}$ been taken.

The formal definitions of ITE (or Individual Causal Effect (ICE)), based on structural causal models, are given in [Pea09, §3]. However, these definitions are not necessary for understanding the Probabilities of Causation (PoCs) in this dissertation. A "unit" means any entity (e.g., a patient, a customer, or an agricultural plot) whose behavior affects decisions. Beyond its measured features $C(u)$, u contains *all* characteristics of an individual u, measured and unmeasured, known and unknown, sufficiently detailed to make the response Y a deterministic function of the treatment. In contrast, population-based decision making optimizes the CATE in Definition 3.1.2.

Definition 3.1.2 (Conditional Average Treatment Effect (CATE))**.**

$$CATE(u) \triangleq E[Y(1,u') - Y(0,u')|C(u') = C(u)] \tag{3.2}$$

where $C(u)$ represents observed pretreatment characteristics of individual u, and the average is taken over all units u' that share these characteristics.

Theoretically, CATE can be viewed as a function of c, the vector of pretreatment characteristics observed on individual u, since CATE will be equal for two different u and u' such that $C(u) = C(u') = c$. However, in order to emphasize the distinction between u, the individual for whom a decision is contemplated, and the individuals u' in the study, CATE(u) = CATE($C(u) = c$) is explicitly written. Clearly, an individual u with characteristics $C(u) = c$ obtains a unique CATE measure, given by Equation (3.2).

This chapter shows that the two objective functions, ITE and CATE, lead to different decision strategies and that, although ITE(u) is in general not identifiable, informative bounds on the probability distribution of ITE(u), for any given individual u, can nevertheless be obtained from aggregate data by combining experimental and observational studies. Formally, bounds will be derived and explained for the proportion of individuals having particular values for ITE (e.g., $P(\text{ITE}(u) = t)$). When Y is binary, bounds will be analyzed for the following values:

- $P(\text{ITE}(u) = 1) = P(Y(1,u) > Y(0,u))$: the proportion of individuals benefiting from treatment,
- $P(\text{ITE}(u) = -1) = P(Y(1,u) < P(Y(0,u))$: the proportion of individuals harmed by treatment,
- $P(\text{ITE}(u) = -1) = P(Y(1,u) = P(Y(0,u))$: the proportion of individuals either immune from or doomed regardless of treatment.

In this way, the information usually provided by the RCT, ATE(u), is supplemented with two additional parameters that may be crucial for individual decision making. Notably, these

aggregate data allow us to improve decision making at the individual level. These bounds can improve decisions that would otherwise be taken using CATE(u) alone as an objective function.

Both ITE(u) and CATE(u) are properties of an individual u and are defined in terms of counterfactual expressions. However, ITE(u) is in general non-identifiable, while CATE(u) is a *do*-expression:

$$\text{CATE}(u) = E[Y|\text{do}(X = 1), C(u)] - E[Y|\text{do}(X = 0), C(u)]. \tag{3.3}$$

Equation (3.3) is estimable directly from experimental data without invoking counterfactual assumptions. Note that the words "individual" and "personalized" used in this chapter refer to the individual for whom a decision is contemplated and, although ITE(u) is the same for all individuals u' sharing measured characteristics $C(u)$ with u, decisions should vary from u to u' depending on their distinct, often unmeasured, personal utilities and beliefs. The aim here is to inform decision makers of the likely behavior of a randomly chosen individual in the population, so as to match decisions to the distinct preferences and beliefs of individuals u and u'.

For conceptual clarity, well-designed RCTs and observational studies are assumed throughout this dissertation. As such, RCTs are considered as having 100% compliance and no selection bias or any other imperfections that often plague them (e.g., placebo effects). Similarly, observational studies are assumed to provide unbiased estimates of the statistical associations or conditional expectations they are designed to assess.

Trialists are usually suspicious of observational studies because they are either bias-prone or rely on subjective assumptions of "no confounding," which are hardly testable. There are two reasons not to worry. First, this analysis makes no modeling assumptions whatsoever when interpreting observational studies, and second, the presence of confounding in the observational studies actually helps narrow bounds on PoCs.

3.2 Qualitative Example

The target of analysis in this chapter and throughout this dissertation is an individual response to a given treatment, namely, how an individual would react if given treatment and if denied treatment. Since no individual can be subjected to both treatment and its denial, its response function must be inferred from population data, originating from one or several studies. Therefore, the question is, "To what degree can population data inform us about an individual response?"

First, there are two conceptual hurdles. First, why should population data provide *any* information whatsoever on the individual response, and second, why should non-experimental data add any information (regarding individual response) to what we can learn with an RCT alone? The next simple example will demonstrate both points.

An RCT is conducted and no difference between treatment (drug) and control (placebo) is found. Let us say 10% in both treatment and control groups die, while the rest (90%) survive. This makes us conclude that the drug is ineffective, but also leaves us uncertain between a range of competing models. Let us just consider the following two models:

- Model A: the drug has no effect whatsoever on any individual and
- Model B: the drug saves 10% of the population and kills another 10%.

From a policy maker viewpoint, the two models may be deemed equivalent; the drug has zero average effect on the target population. But from an individual viewpoint, the two models differ substantially in the sets of risks and opportunities they offer. According to A, the drug is useless but safe. According to B, however, the drug may be deemed dangerous by some and a life-saver by others.

Greenland and Robins [GR86] named four types of individuals with binary treatment and outcome: doomed, exposure causative, exposure preventative, and immune. To be more in line with the research presented in this dissertation and the expansions on these PoCs

in Chapter 9, I have named the probabilities corresponding to these four response types as follows:

- P(benefit): probability that treatment will benefit an individual,
- P(harm): probability that treatment will harm an individual,
- P(immunity): probability that an individual is immune, regardless of treatment,
- P(doom): probability that an individual is doomed, regardless of treatment.

Models A and B can now be classified according to these PoCs, as shown in Table 3.1.

	Model A	Model B
P(benefit)	0	0.1
P(harm)	0	0.1
P(immunity)	0.9	0.8
P(doom)	0.1	0

Table 3.1: PoCs for models A and B

To see how attitudes about counterfactual effects may emerge, assume, for the sake of argument, that the drug also provides temporary pain relief. Model A would be deemed desirable and safe by all, whereas model B will scare away those who do not urgently need the pain relief, while offering a glimpse of hope to those whose suffering has become unbearable and who would be ready to risk death for the chance (10%) of recovery (hoping, of course, they are among the lucky beneficiaries).

Another reason for diverse individual decisions in the face of B is individual beliefs. For example, a person may believe the drug will not be harmful to them even though it has a 10% probability of harm in the population at large. Maybe a family member took the drug and recovered. In that case, the drug certainly was not harmful to that family member. This

person would choose to take the drug under B (the drug saves 10% of the population and kills another 10%) and reject it under A because, assuming immunity, there are still factors of pain, expense, and discomfort to consider. A different person, whose cousin happened to die after taking the drug, may not have confidence in natural immunity and will choose to refuse the drug under B.

It should now be clear that individuals can have unique preferences for how they weigh the probability of benefit and the probability of harm. They may additionally place weights on the probability of being doomed (death, regardless of taking the drug or not) and the probability of being immune (recovery, regardless of taking the drug or not). Li and Pearl [LP19] provide bounds, or point estimates when certain assumptions can be made, on a linear combination of the probabilities of benefit, harm, immunity, and doom. They demonstrate how the ATE is a sub-optimal criterion for decision making in light of the weights pertaining to individual preferences. Mueller and Pearl [MP20] supply an example of this weighting of probability of benefit and probability of harm in assessing which Covid-19 patients are in greatest need of treatment.

This simple example will also serve to illustrate the crucial role of observational studies. Supplementing the RCT with an observational study on the same population (conducted, for example, by an independent survey of patients who have the option of taking or avoiding the drug) would allow a determination between the two models, totally changing the understanding of what risks await an individual taking the drug.

Consider an extreme case where the observational study shows 100% survival in both drug-choosing and drug-avoiding patients, as if each patient knew in advance where danger lies and managed to avoid it. Such a finding, though extreme and unlikely, immediately rules out A which claims no treatment effect on any individual. This is because the mere fact that patients succeed 100% of the time to avoid harm where harm does exist (revealed through the 10% death in the randomized trial) means that choice makes a difference, contrary to A's claim that choice makes no difference.

The same argument applies when the probability of survival among option-having individuals is not precisely 100% but simply higher (or lower) than the probability of survival in the RCT. Using the RCT study alone, in contrast, we were unable to rule out A, or even to distinguish A from B.

Consider another edge case where B, rather than A, is ruled out as impossible. Assume the observational study informs us that all those who chose the drug died and all who avoided the drug survived. It seems that drug-choosers were unfortunately bad decision makers while drug-avoiders knew precisely what was good for them. This is perfectly feasible, one possibility is that 10% of people who would be doomed regardless of their drug choice chose to take the drug (not knowing that they are doomed anyway and could have saved themselves the cost, pain, and trouble) and the remaining 90% of people who are immune regardless of their drug choice chose not to take the drug (smart). However, this also tells us that no one can be *cured* by the drug, contrary to the assertion made by B, that the drug cures 10% and kills 10%. To be cured, a person must survive if treated and die if not treated. But none of the drug-choosers could have been cured, because they all died, and none of the drug avoiders could have been cured, because they all survived (they might have survived had they taken the drug, but then it would not have been the drug that cured them). Thus, B cannot explain these observational results, and must be ruled out.

Although B tells an individual that she has a 10% chance of being killed by the drug and a 10% chance of being saved by the drug, she cannot know which outcome pertains to her specifically. All she knows is that 10% of people with her characteristics are killed by the drug and another 10% will be cured due to the drug. This is information that is suppressed by the RCT which does not distinguish between A and B.

Now that it has been demonstrated conceptually how certain combinations of observational and experimental data can provide information on individual behavior that each study alone cannot, the next step is a more realistic motivating example which, based on theoretical bounds derived in [TP00], establishes individual behavior for any combination of

observational and experimental data and, moreover, demonstrates critical decision making ramifications of the information obtained. Note that the example below happened to be identifiable due to particular combinations of data, though, in general, the data will not permit point estimates of individual causal effects and the bounds will not necessarily be narrow. This motivating example is merely an extreme case concocted to explain the origin of the bound-narrowing effect.

3.3 Motivating Numerical Example

Consider the effect of a drug on two subpopulations, males and females. Unlike the extreme case considered in Section 3.2, the drug is found to be somewhat effective for both males and females and, in addition, deaths are found to occur in the observational study as well. Although men and women are totally indistinguishable in the RCT study, adding observational data proves men to react markedly different than women, calling for two different treatment policies in the two groups. Whereas a woman has a 28% chance of benefiting from the drug and no danger at all of being harmed by it, a man has a 49% chance of benefiting from it and as much as a 21% chance of dying because of it.

To cast the story in a realistic setting, imagine the testing of a new drug, aimed to help patients suffering from a deadly disease. An RCT is conducted to evaluate the efficacy of the drug and it is found to be 28% effective in both males and females; in other words, CATE(male) = CATE(female) = 0.28. To simplify matters, assume each experimental study data as an ideal RCT, with 100% compliance and no selection bias or any other biases that often plague RCTs.

The RCT tells us that there was a 28% improvement, on average, in taking the drug compared to not taking the drug. This was the case among both females and males: CATE(female) = CATE(male) = 0.28, where do(drug) and do(no-drug) are the treatment and control arms in the RCT. It thus appears reasonable to conclude that the drug has a

	Female Survivals	Male Survivals
do(drug)	489/1000 (49%)	490/1000 (49%)
do(no drug)	210/1000 (21%)	210/1000 (21%)
CATE	28%	28%

Table 3.2: Female versus male CATE

net remedial effect on some patients and that every patient, be it male or female, should be advised to take the drug and benefit from its promise of increasing one's chances of recovery by 28%.

At this point, the drug manufacturer ventured to find out to what degree people actually buy the approved drug, following its recommended usage. A market survey was conducted (observational study) and revealed that only 70% of men and 70% of women actually chose to take the drug; problems with side effects and rumors of unexpected deaths may have caused the other 30% to avoid it. As with the experimental studies, observational studies are assumed to provide unbiased estimates for the conditional probabilities involved. Note that observational studies provide an easier arena for obtaining representative samples of the target population, partly due to the ease of recruiting units and partly due to their non-invasive nature. A careful examination of the observational study has further revealed substantial differences in survival rates of men and women who chose to use the drug (shown in Tables 3.3 and 3.4). The rate of recovery among drug-choosing men was exactly the same as that among the drug-avoiding men (70% for each), but the rate of recovery among drug-choosing women was 43% lower than among drug-avoiding women (0.27 versus 0.70, in Table 3.3). It appears as though many women who chose the drug were already in an advanced stage of the disease, which may account for their low recovery rate of 27%.

At this point, having data from both experimental and observational studies we can estimate the probability $P(\text{ITE}(u) > 0 | C(u) = c) = P(\text{benefit} | C(u) = c)$ for both a typical man and a typical woman. Quantitative analysis shows (see Section 3.5) that, with the data

		Survivals	Deaths	Total
Experimental	*do*(drug)	489 (49%)	511 (51%)	1,000 (50%)
	do(no drug)	210 (21%)	790 (79%)	1,000 (50%)
Observational	drug	378 (27%)	1,022 (73%)	1,400 (70%)
	no drug	420 (70%)	180 (30%)	600 (30%)

Table 3.3: Female survival and recovery data

		Survivals	Deaths	Total
Experimental	*do*(drug)	490 (49%)	510 (51%)	1,000 (50%)
	do(no drug)	210 (21%)	790 (79%)	1,000 (50%)
Observational	drug	980 (70%)	420 (30%)	1,400 (70%)
	no drug	420 (70%)	180 (30%)	600 (30%)

Table 3.4: Male survival and recovery data

above, the drug affects men markedly differently from the way it affects women. Whereas a woman has a 28% chance of benefiting from the drug and no danger at all of being harmed by it, a man has a 49% chance of benefiting from it and as much as a 21% chance of dying because of it — a serious cause for concern. Note that based on the experimental data alone (Table 3.2), no difference at all can be noticed between men and women.

The ramifications of these findings on personal decision making are broad. First, they tell us that the drug is not as safe as the RCT would have us believe; it may cause death in a sizable fraction of patients. Second, they tell us that a woman is totally clear of such dangers, and should have no hesitation to take the drug, unlike a man, who faces a decision; a 21% chance of being harmed by the drug is cause for concern. Physicians, likewise, should be aware of the risks involved before recommending the drug to a man. Third, the data tell policy makers what the overall societal benefit would be if the drug is administered to women only; 28% of the drug-takers would survive who would otherwise die. Finally, knowing the

relative sizes of the benefiting versus harmed subpopulations opens the door to finding the mechanisms responsible for the differences, as well as to identifying measurable markers that characterize those subpopulations.

Additional measured features, other than gender, can be leveraged, such as family history, a genetic marker, or a side-effect, and checked whether they shrink the sizes of susceptible subpopulations. The results would be a set of features that approximate responses at the individual level. Note that absent observational data and a calculus for combining them with the RCT data, we would not be able to identify such informative features. A feature like gender would be deemed irrelevant, since men and women were indistinguishable in the RCT studies.

The ability to identify relevant informative features can be leveraged to amplify the potential benefits of the drug. For example, if we identify a marker that characterizes men who would die only if they take the drug and prevent those patients from taking the drug, the drug would cure 62% of male patients who would be allowed to use it. This is because we would not administer the drug to the 21% who would have been killed by the drug. Those patients will now survive, so a total of 70% of patients will be cured because of this combination of marker identification and drug administration. This unveils an enormous potential of the drug at hand, which was totally concealed by the 28% effectiveness estimated in the RCT studies.

3.4 Notation

The following notational conventions will be adopted for the remainder of this dissertation until an expansion is necessary for Chapter 9 or when otherwise specified. Let random variable X represent a binary treatment, with x = true and x' = false. Similarly, let the random variable Y represent binary outcome, with y = true and y' = false. In clinical study settings, the analogous might be assigned: x = treated, x' = untreated, y = recovered, and

y' = unrecovered.

The counterfactual notation used in Pearl's *Causality* [Pea09] will be adopted. $Y_x = y$ denotes the counterfactual sentence, "Variable Y would have the value y, had X been x." This event is further simplified with the notation y_x, such that $P(Y_x = y) \triangleq P(y_x)$. There are four probabilities of this form with binary treatment and binary outcome: $P(y_x)$, $P(y_{x'})$, $P(y'_x)$, and $P(y'_{x'})$.

3.5 How the Results Were Obtained

Let us denote the outcome variable Y as recovery and the treatment variable X as treated. The causal effects for treatment and control groups, $P(y_x|\text{Gender})$ and $P(y_{x'}|\text{Gender})$, were the same. No differences were noted between males and females. Note that $P(y_x|\text{female})$ was rounded up from 48.9% to 49%. The 0.001 difference between $P(y_x|\text{female})$ and $P(y_x|\text{male})$ wasn't necessary, but was constructed to allow for clean point estimates.

In addition to the above RCT, an observational study (survey) was conducted on the same population. Let us denote $P(y|x, \text{Gender})$ and $P(y|x', \text{Gender})$ as recovery among the drug-choosers and recovery among the drug-avoiders, respectively.

With this notation at hand, let us define the probability of benefit.

Definition 3.5.1 (Probablity of Benefit (P(benefit)))**.**

$$P(\textit{benefit}) \triangleq P(y_x, y'_{x'}) \tag{3.4}$$

The probability P(benefit) will be computed from the following data sources: $P(y_x)$, $P(y_{x'})$, $P(y|x)$, $P(y|x')$, and $P(x)$. The first two denote the data obtained from the RCT and the last three, data obtained from the survey. Non-recovery is represented by y', so $y'_{x'}$ is non-recovery among the RCT control group. Equation (3.4) should be interpreted as the probability that an individual would both recover if assigned to the RCT treatment arm and die if assigned to control. Tian and Pearl [TP00] called P(benefit), "Probability of Necessity

and Sufficiency" (PNS).

The results of the observational and experimental studies are not independent of each other since, barring selection bias, participants in the two studies are selected from the same overall population, ideally consisting of the eventual users of the drug. At the individual level, the connection between behaviors in the two studies relies on an assumption known as *consistency* [Pea09, Pea10], asserting that an individual's response to treatment depends entirely on biological factors, unaffected by the settings in which treatment is chosen. In other words, the outcome of a person choosing the drug would be the same had this person been assigned to the treatment group in an RCT study. Similarly, if we observe someone avoiding the drug, their outcome is the same as if they were in the control group of our RCT.

Consistency is a property imposed at the individual level, often written as

$$Y = X \cdot Y(1) + (1 - X) \cdot Y(0)$$

for binary X and Y. Rubin [Rub74] considered consistency to be an assumption in SUTVA, which defines the potential outcome (PO) framework. Pearl [Pea10] considered consistency to be a theorem of Structural Equation Models (SCMs), a violation of which reflects imperfections (e.g., placebo effects) in RCT practices.

In medical practices, clinical experts rarely rely on the assumption of biological equivalence. The very participation in a study tends to create fears and expectations that affect patients' response to treatment. Moreover, selection bias [BTP14] is a major problem in clinical trials, since subjects are recruited by stringent health criteria and, unlike those in observational studies, they must undergo consent procedures. For these two reasons, RCT practitioners compare only patients that undergo the same recruitment procedure and, accordingly, report only the difference $P(y_x) - P(y_{x'})$. More elaborate procedures [BTP14] must be deployed to overcome both selection bias and placebo effects when experimental and observational studies are to be combined.

Consistency implies:

$$P(y_x|x) = P(y|x), P(y_{x'}|x') = P(y|x'). \tag{3.5}$$

In words, the probability that a drug-chooser would recover in the treatment arm of the RCT, $P(y_x|x)$, is the same as the probability of recovery in the observational study, $P(y|x)$.

Based on this assumption, and leveraging both experimental and observational data, Tian and Pearl [TP00] derived the following tight bounds on the probability of benefit, as defined in (3.4):

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \leqslant P(\text{benefit}) \leqslant \min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) \\ + P(x,y') + P(x',y) \end{array} \right\}. \tag{3.6}$$

Here $P(y'_{x'})$ is equivalent to $1 - P(y_{x'})$, namely the probability of death in the control group. The same bounds hold for any subpopulation, say males or females, if every term in (3.6) is conditioned on the appropriate class.

Applying these expressions to the female data from Table 3.3 gives the following bounds on $P(\text{benefit}|\text{female})$:

$$\begin{aligned} \max\{0, 0.279, 0.09, 0.189\} \leqslant P(\text{benefit}|\text{female}) \leqslant \min\{0.489, 0.79, 0.279, 1\}, \\ 0.279 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.279. \end{aligned} \tag{3.7}$$

Similarly, for men we get:

$$\begin{aligned} \max\{0, 0.28, 0.49, -0.21\} \leqslant P(\text{benefit}|\text{male}) \leqslant \min\{0.49, 0.79, 0.58, 0.7\}, \\ 0.49 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.49. \end{aligned} \tag{3.8}$$

Thus, the bounds for both females and males, in (3.7) and (3.8), collapse to point estimates:

$$P(\text{benefit}|\text{female}) = 0.279,$$
$$P(\text{benefit}|\text{male}) = 0.49.$$

We aren't always so fortunate to have a complete set of observational and experimental data at our disposal. When some data is absent, we are allowed to discard arguments to max or min in (3.6) that depend on that data. For example, if we lack all experimental data, the only applicable lower bound in (3.6) is 0 and the only applicable upper bound is $P(x, y) + P(x', y')$:

$$0 \leqslant P(\text{benefit}) \leqslant P(x, y) + P(x', y'). \tag{3.9}$$

Applying these observational data only bounds to males and females yields:

$$0 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.279,$$
$$0 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.58.$$

Naturally, these are far more loose than the point estimates when combined experimental and observational data is fully available. Let's similarly examine what can be computed with purely experimental data. Without observational data, only the first two arguments to max of the lower bound and min of the upper bound of $P(\text{benefit})$ in (3.6) are applicable:

$$\max\{0, P(y_x) - P(y_{x'})\} \leqslant P(\text{benefit}) \leqslant \min\{P(y_x), P(y'_{x'})\}. \tag{3.10}$$

Applying these bounds (using only experimental data) to males and females yields:

$$0.279 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.489,$$
$$0.28 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.49.$$

Again, these are fairly loose bounds, especially when compared to the point estimates obtained with combined data. Notice that the overlap between the female bounds using observational data, $0 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.279$, and the female bounds using experimental

data, $0.279 \leqslant P(\text{benefit}|\text{female}) \leqslant 0.489$ is the point estimate $P(\text{benefit}|\text{female}) = 0.279$. The more comprehensive Tian-Pearl bounds formula (3.6) wasn't necessary. However, the intersection of the male bounds using observational data, $0 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.58$, and the male bounds using experimental data, $0.28 \leqslant P(\text{benefit}|\text{male}) \leqslant 0.49$, does not provide us with narrower bounds beyond what the experimental data provides. For males, the comprehensive Tian-Pearl bounds in (3.6) was necessary for narrower bounds (in this case, a point estimate).

Having seen this mechanism of combining observational and experimental data in (3.6) work so well, it is natural to ask, "what's behind this?" The intuition comes from the fact that observational data incorporates individuals' whims, and whims are proxies for hidden factors that may affect that individual's response to treatments. Such "confounding" factors are usually problematic in causal inference, since they lead to biased conclusions, sometimes completely reversing a treatment's effect [Pea13]. Confounding then needs to be adjusted for. However, here confounding helps us, exposing the underlying mechanisms its associated whims and desires are a proxy for.

Finally, as noted in Section 3.3, knowing the relative sizes of the benefiting versus harmed subpopulations demands investment in finding mechanisms responsible for the differences as well as characterizations of those subpopulations. For example, women above a certain age might be affected differently by the drug, which could be detected by investigating how age affects the bounds on the individual response. Such characteristics can potentially be narrowed repeatedly until the drug's efficacy can be predicted for an individual with certainty or the underlying mechanisms of the drug can be fully understood.

None of this was possible with only the RCT. Yet, remarkably, an observational study, however sloppy and uncontrolled, provides a deeper perspective on a treatment's effectiveness. It incorporates individuals' whims and desires that govern behavior under free-choice settings. And, since such whims and desires are often proxies for factors that also affect outcomes and treatments (i.e., confounders), we gain additional insight hidden by RCTs.

3.6 STAR Real World Example

As a real world example, let us consider STAR project from Chapter 2 with binary discretized outcomes. The experimental and observational probabilities are:

$$P(y_x) = 0.7914 \qquad P(y'_x) = 0.2086, \tag{3.11}$$

$$P(y_{x'}) = 0.7460 \qquad P(y'_{x'}) = 0.2540, \tag{3.12}$$

$$P(y|x) = 0.7789 \qquad P(y'|x) = 0.2211, \tag{3.13}$$

$$P(y|x') = 0.8947 \qquad P(y'|x') = 0.1053, \tag{3.14}$$

$$P(x) = 0.1925 \qquad P(x') = 0.8075, \tag{3.15}$$

where y is a high second grade math score, y' is a low second grade math score, x is a small second grade class size, and x' is a regular second grade class size.

We can now estimate $P(\text{benefit})$ from Equation (3.6):

$$\begin{aligned} \max\{0, 0.0454, 0.1264, -0.081\} \leqslant P(\text{benefit}) \leqslant \min\{0.7914, 0.254, 0.235, 0.8104\}, \\ 0.1264 \leqslant P(\text{benefit}) \leqslant 0.235, \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} P(y) &= P(y|x) \cdot P(x) + P(y|x') \cdot P(x) \\ &\approx 0.8724. \end{aligned}$$

It is interesting to note that the ATE for how small class sizes affect math performance is $P(y_x) - P(y_{x'}) = 0.0454$. This is notably smaller than even the lower bound of $P(\text{benefit})$ in Equation (3.16). The Tennessee State Department of Education may have determined that small classes do have a small positive effect. However, they may have increased funding for this project, and other states may have taken notice, had they known that at least 12.64% up to almost a quarter of students benefited.

Furthermore, combining both experimental and observational data was necessary to show how much better students benefited than the ATE indicated. By only taking into account observational data, the lower bound of $P(\text{benefit})$ is 0:

$$0 \leqslant P(\text{benefit}) \leqslant 0.235.$$

While only taking into account experimental data, the lower bound is the ATE:

$$0.0454 \leqslant P(\text{benefit}) \leqslant 0.235.$$

3.7 Probability of Benefit Intuition

The Tian-Pearl bounds on $P(\text{benefit})$ in Equation (3.6) were discovered through linear programming using constraints on observational probabilities, interventional probabilities, and consistency. It is instructive to dissect these bounds and understand why they hold and why they are tight. This will provide necessary background for the rest of this dissertation where the tightness of these bounds is overcome with reasonable assumptions and then these bounds are generalized beyond binary outcomes.

Lower Bound

The first two arguments to max in Equation (3.6) come from Fréchet Inequalities [MP19] for two events:

$$\max(0, P(A) + P(B) - 1) \leqslant P(A, B) \leqslant \min(P(A), P(B)) \tag{3.17}$$

where $A = y_x$ and $B = y'_{x'}$.

Let us now assume, w.l.o.g., that x represents treatment and x' represents no treatment, while y represents successful outcome and y' represents unsuccessful outcome.

For the third argument to the max function, $P(y) - P(y_{x'})$, we can imagine the individuals represented by y and the individuals represented by $y_{x'}$. The outcome $Y = y$ (successful

outcome in an observational study) consists of some individuals who benefit from treatment (the ones who chose treatment), some individuals who are harmed by treatment (the ones who chose no treatment), and all individuals who have a positive outcome regardless of treatment (the immune). The outcome $Y_{x'} = y$ (successful outcome had no treatment been administered) consists of all individuals who are harmed by treatment $(y_{x'}, y'_x)$ and all individuals who have a positive outcome regardless of treatment (the immune, $y_{x'}, y_x$). Therefore,

$$\begin{aligned} P(y) - P(y_{x'}) &= P(\text{some benefiters} + \text{some harmed} + \text{all immune}) \\ &\quad - P(\text{all harmed} + \text{all immune}) \\ &= P(\text{some benefiters}) + [P(\text{some harmed}) - P(\text{all harmed})] \\ &= (\text{number } \leqslant P(\text{benefit})) + \text{non-positive number} \\ &\leqslant P(\text{benefit}). \end{aligned}$$

Similarly, for the fourth argument to the max function

$$\begin{aligned} P(y_x) - P(y) &= P(\text{all benefiters} + \text{all immune}) \\ &\quad - P(\text{some benefiters} + \text{some harmed} + \text{all immune}) \\ &= [P(\text{all benefiters}) - P(\text{some benefiters})] - P(\text{some harmed}) \\ &= \text{number less than } P(\text{benefit}) - \text{non-negative number} \\ &\leqslant P(\text{benefit}). \end{aligned}$$

Upper Bound

Just like the lower bound, the first two arguments to min come from Fréchet Inequalities, where $A = y_x$ and $B = y'_{x'}$.

For the third argument to the min function, $P(x, y) + P(x', y')$, we can dissect these probabilities similar to how we dissected the third argument to the max function. The probability $P(x, y)$ is the proportion of people who chose treatment and had a successful

outcome, which includes benefiters who chose treatment and those immune to the negative outcome who chose treatment. The probability $P(x', y')$ is the proportion of people who chose no treatment and had an unsuccessful outcome, which includes benefiters who chose no treatment and those doomed to the negative outcome who chose no treatment. Note that the other two possible scenarios, (x, y') and (x', y), cannot contain any benefiters. Therefore,

$$\begin{aligned} P(x,y) + P(x',y') &= P(\text{benefiters choosing treatment} + \text{some immune}) \\ &\quad + P(\text{benefiters choosing no treatment} + \text{some harmed}) \\ &= P(\text{benefit}) + [P(\text{some immune}) + P(\text{some harmed})] \\ &= P(\text{benefit}) + \text{non-negative number} \\ &\leqslant P(\text{benefit}). \end{aligned} \tag{3.18}$$

For the fourth argument to the min function, $P(y_x) - P(y_{x'}) + P(x, y') + P(x', y)$, let us rewrite the expression to aid intuition:

$$\begin{aligned} P(\text{benefit}) &\leqslant P(y_x) - P(y_{x'}) + P(x,y') + P(x',y) \\ &= P(y_x) - [1 - P(y'_{x'})] + (1 - [P(x,y) + P(x',y')]) \\ &= P(y_x) + P(y'_{x'}) - [P(x,y) + P(x',y')]. \end{aligned}$$

Now we can use the fact that y_x represents all benefiters and all immune individuals, while $y'_{x'}$ represents all benefiters and all doomed individuals. Combining that with Equation (3.18),

$$\begin{aligned} P(\text{benefit}) &\leqslant P(\text{benefiters} + \text{immune}) + P(\text{benefiters} + \text{doomed}) \\ &\quad - P(\text{benefiters} + \text{some immune} + \text{some harmed}) \\ &= P(\text{benefit}) + [P(\text{some immune}) + P(\text{some harmed})] \\ &= P(\text{benefit}) + \text{non-negative number} \\ &\leqslant P(\text{benefit}). \end{aligned}$$

3.8 ATE and Probabilities of Harm, Immunity, and Doom

It may seem like P(benefit) is the true probability that matters in decision making and counterfactual reasoning. However, as will be shown in Chapters 6, 7, and 9, the probabilities of harm, immunity, and doom are critical to more precise and narrow bounds on results of interest and personalized decision making.

The bounds on these remaining PoCs can be derived through linear programming, algebraically, or qualitatively as above. However, there's an easier approach. First, let us define these remaining PoCs formally.

Definition 3.8.1 (Probablity of Harm (P(harm)))**.**

$$P(harm) \triangleq P(y'_x, y_{x'}) \tag{3.19}$$

Definition 3.8.2 (Probablity of Immunity (P(immunity)))**.**

$$P(immunity) \triangleq P(y_x, y_{x'}) \tag{3.20}$$

Definition 3.8.3 (Probablity of Doom (P(doom)))**.**

$$P(doom) \triangleq P(y'_x, y'_{x'}) \tag{3.21}$$

The probabilities P(benefit) and P(harm) are related to the ATE:

$$\begin{aligned} ATE &\triangleq E[Y_x - Y_{x'}] \\ &= P(y_x) - P(y_{x'}) \\ &= [P(y_x, y_{x'}) + P(y_x, y'_{x'})] - [P(y_{x'}, y_x) + P(y_{x'}, y'_x)] \\ &= P(y_x, y_{x'}) - P(y_{x'}, y'_x) \\ &= P(\text{benefit}) - P(\text{harm}). \end{aligned} \tag{3.22}$$

Note that the second equality only applies when the outcome variable, Y, is binary. In that case, it can always be transformed to take on the values $Y = y = 1$ and $Y = y' = 0$. This will significantly affect our derivations in Chapter 9, when we consider non-binary outcomes.

Equation (3.22) helps in two ways. First, it tells us immediately that when no harm is possible (i.e., $P(\text{harm}) = 0$), $P(\text{benefit})$ coincides with ATE, or, in other words, ATE constitutes a point estimate of $P(\text{benefit})$. The concept of zero harm is known as monotonicity and will play a significant role in Chapters 7 and 9. Second, it allows us to compute $P(\text{harm})$ from $P(\text{benefit})$ and ATE in cases where monotonicity does not hold, as was the case for men in the numeric example of Section 3.3.

For each of females and males, in the above example, their respective $P(\text{benefit})$ and ATE are known. Therefore their probabilities of harm are known as well:

$$\begin{aligned} P(\text{harm}|\text{female}) &= P(\text{benefit}|\text{female}) - \text{CATE}(\text{female}) \\ &= 0.279 - 0.279 = 0, \\ P(\text{harm}|\text{male}) &= P(\text{benefit}|\text{male}) - \text{CATE}(\text{male}) \\ &= 0.49 - 0.28 = 0.21. \end{aligned}$$

Deriving $P(\text{harm})$ involves simply subtracting $P(y_x) - P(y_{x'})$ from each argument of the min and max functions of Equation (3.6):

$$\max \left\{ \begin{array}{c} 0, \\ P(y_{x'}) - P(y_x), \\ P(y) - P(y_x), \\ P(y_{x'}) - P(y) \end{array} \right\} \leqslant P(\text{harm}) \leqslant \min \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x, y') + P(x', y), \\ P(y_{x'}) - P(y_x) \\ + P(x, y) + P(x', y') \end{array} \right\}. \tag{3.23}$$

The astute reader may have noticed that simply swapping x for x' and vice-versa from Equation (3.6) would have sufficed.

The probabilities of immunity and doom can similarly be derived through their relationships with $P(\text{benefit})$ and $P(\text{harm})$:

$$P(\text{immunity}) = P(y_x) - P(\text{benefit}), \tag{3.24}$$

$$P(\text{doom}) = P(y'_x) - P(\text{harm}), \tag{3.25}$$

The bounds on $P(\text{immunity})$ and $P(\text{immunity})$ are then computed straightforwardly from Equations (3.24) and (3.25):

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y'_{x'}), \\ P(y_x) - P(x,y) - P(x',y') \\ P(y_{x'}) - P(x,y') - P(x',y) \end{array} \right\} \leqslant P(\text{immunity}) \leqslant \min \left\{ \begin{array}{c} P(y_x), \\ P(y_{x'}), \\ P(y_x) + P(y_{x'}) - P(y), \\ P(y) \end{array} \right\}, \tag{3.26}$$

$$\max \left\{ \begin{array}{c} 0, \\ P(y'_x) - P(y_{x'}), \\ P(y'_x) - P(x,y') - P(x',y) \\ P(y'_{x'}) - P(x,y) - P(x',y') \end{array} \right\} \leqslant P(\text{doom}) \leqslant \min \left\{ \begin{array}{c} P(y'_x), \\ P(y'_{x'}), \\ P(y'_x) + P(y'_{x'}) - P(y'), \\ P(y') \end{array} \right\}. \tag{3.27}$$

Notice that deriving $P(\text{doom})$ from $P(\text{immunity})$ and vice-versa simply requires swapping y for y' and vice-versa.

3.8.1 Visualization

I developed a visualization of $P(\text{harm})$ and the Tian-Pearl bounds of $P(\text{benefit})$ at `https://learn.ci/bounds.html`, shown in Figure 3.1, in order to allow users to gain an intuition around bounds on probabilities of causation. Currently this visualization does not visualize other PoCs, nor does it incorporate covariates, mediators, ATE, intention or treatment evidence (presented in Chapter 6), or other mechanisms in upcoming chapters to narrow or modify the bounds. Future visualizations should incorporate some, all of, or more than these capabilities.

Note that the top left of the visualization, as well as the hover box upon clicking "Display data when hovering", displays $P(Y_x > Y_{x'})$ for $P(\text{benefit})$ instead of the $P(y_x, y_{x'})$ discussed in this and the previous chapters. The reason for this alternative notation is to account for

non-binary ordinal outcomes. This will be presented in Chapter 9. Currently this visualization only supports binary treatment and outcomes. Future visualizations should support at least non-binary ordinal outcomes. Visualizing non-binary ordinal treatments is a bit more challenging and will require some clever user interface and graphical mechanisms.

An additional aspect of this visualization to notice is that there is a "Possible region" window. Any points outside of this window are incompatible. This means that the interventional probability values of $P(y_x)$ and $P(y_{x'})$ that make up a coordinate outside the possible region cannot occur while the observational probability values of $P(x)$, $P(y|x)$, and $P(y|x')$ are set to their specific values through their associated sliders. For example, it is impossible for the probability of success had treatment occurred to be 100% ($P(y_x) = 1$) while nobody succeeded among people who chose treatment ($P(y|x) = 0$) and everyone chose treatment ($P(x) = 1$).

3.9 Summary

Most causal inference research has focused on the ATE, a population quantity that is estimable directly from RCTs, but provides no information on how individuals who respond one way under treatment would respond under an alternative. Theoretical results in this dissertation show that we can go beyond ATE, to estimate (or bound) the entire distribution of individual causal effects. The bounds estimated can be quite narrow and allow us to make accurate personalized decisions. In other words, a randomly chosen individual would be able to assess how she would respond both to treatment and to its negation and, accordingly, decide on a course of action that best fits both her personal preferences and societal needs.

The key to getting accurate information on individual causal effects lies in combining observational and experimental data. While the mathematics of this combination was developed two decades ago [Pea99, TP00], this chapter explains the mechanism behind it, and the rest of this dissertation demonstrates its implications in decision making situations

and updates the math in various ways to get more precise bounds. Observational data reveal individual idiosyncrasies that are masked in experimental settings. In other words, the unobserved confounding factors that usually affect both treatments and outcomes in observational studies are proxies for individual idiosyncrasies, often emanating from unique experience, beliefs, or desires that also govern individual treatment effects.

Having established the mechanism of combining observational and experimental data, this chapter demonstrated the added value of observational studies in situations where experimental data are available. For example, gaining access to observational data may flip individual decisions from treatment to no-treatment and vice versa. From a policy-making viewpoint, information obtained from observational studies may reveal significant heterogeneity among subpopulations (e.g., males versus females) that are totally indistinguishable in experimental settings. Such information can identify which subpopulations are most susceptible to harmful treatment effects and thus assist in differential policy making. Finally, identifying susceptible subpopulations opens the possibility of searching for the mechanisms responsible for their differences as well as identifying new predictive markers associated with those differences.

While these findings illuminate individual responses to treatment and may help individual decision making, they apply equally to all individuals sharing u's measured characteristics, $C(u)$. Conditioning on additional characteristics of the individual involved should provide, of course, additional person-specific information. However, such additions are accompanied with increased variance and must therefore be limited by the sample size available in each stratum. The bounds in this chapter and the rest of this dissertation are not subject to this limitation and takes full advantage of the large sample size usually available in observational studies. These methods provide the key for next-generation personalized decision making.

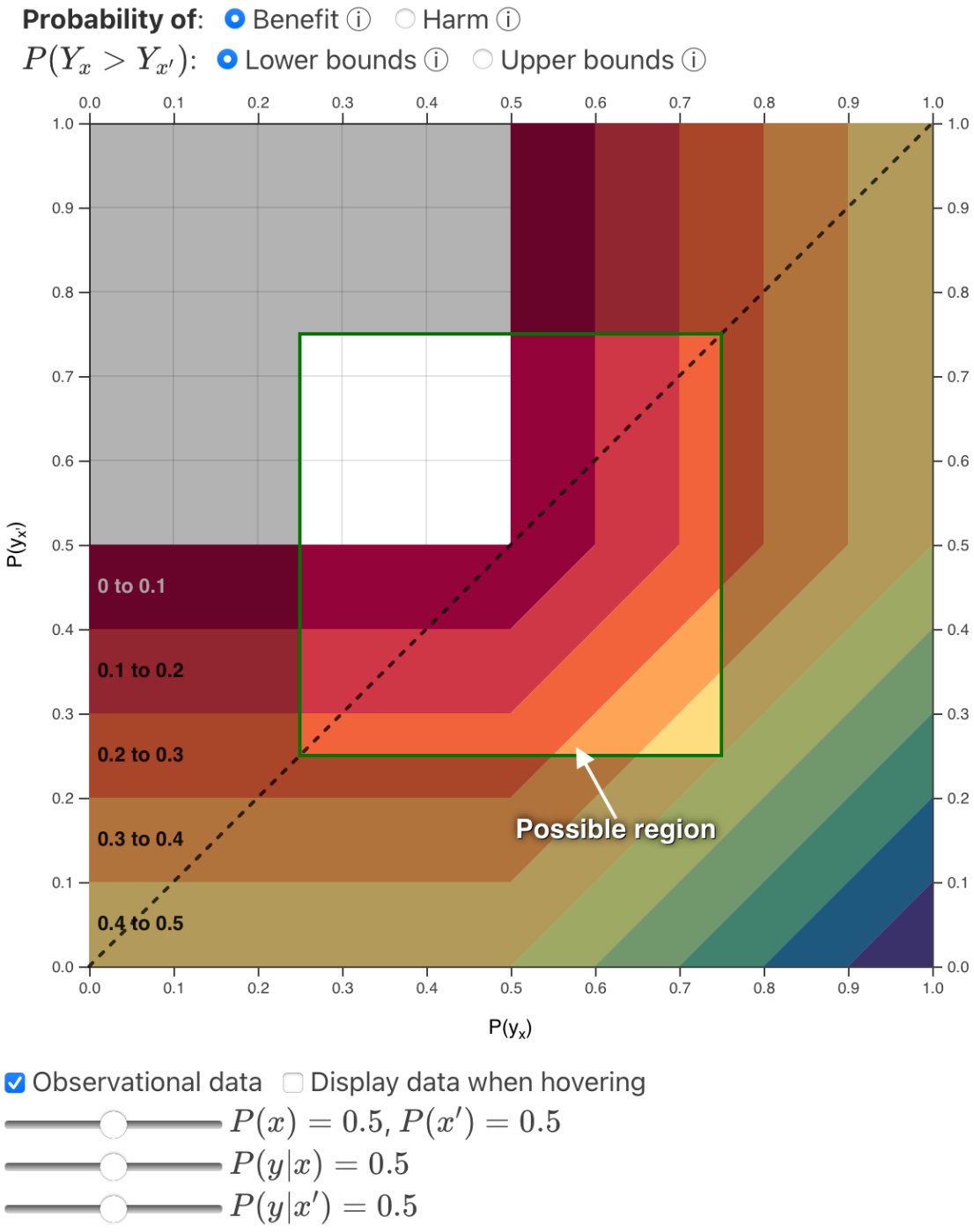


Figure 3.1: Visualization of P(harm) and the Tian-Pearl bounds of P(benefit).

CHAPTER 4

Estimating Probabilities of Causation

4.1 Introduction

Most analyses in the past three decades have been concerned with estimating effects of causes (EoC). Less emphasis has been placed on identifying causes of effects (CoE), despite their critical importance in science, medicine, public policy, legal reasoning, AI, and epidemiology [Pea15].

One reason for this imbalance is that without strong assumptions or knowledge of the underlying functional form, we can generally only obtain bounds, as opposed to point estimates, on PoCs from statistical data. Unfortunately, bounds on these probabilities, such as those in Equations (3.6), (3.23), (3.26), and (3.27), are often too loose to be informative, assumptions necessary for point estimates are often too strong to be applicable, and the functional form is seldom known. These issues may constitute some of the reasons CoE is used and researched far less than EoC.

Learning functions of EoC has been greatly aided by significant advances in machine learning. Huge quantities of very high-dimensional data can be processed for accurate EoC. This enables us to create better policies for a population, such as whether a mRNA vaccine, a protein subunit vaccine, or a vector vaccine is most effective for a particular subpopulation of given age, sex, and other characteristics. However, these results can be surprisingly misleading in the context of personal decision making. A randomized controlled trial (RCT) doesn't eliminate this deception. The following is an example of this scenario that was

presented in [MP20].

Sadly, many regions in the world experienced shortages of SARS-CoV-2 vaccines during times of high infection rates. Ideally, they would administer their limited supply to those most in need. In order to do this, they would need to identify subpopulations with the highest probabilities of *both* surviving if vaccinated and dying if unvaccinated.

A first step is to determine which characteristics or variables are highly correlated with recovery. The analyses in [MP20] and Chapter 3 focused on gender. However, [MLP22] used the following, more realistic, scenario of a machine learning algorithm discovering a high correlation between age and recovery. We classified ages into two groups: under sixty years old and over 59 years old. From an RCT, it was determined that older people have an ATE of 20% (or 0.2), as 57% survive when vaccinated and 37% survive when unvaccinated. These survival rates are artificially low for demonstration purposes, but we can imagine a region with extremely high infection rates for a particularly virulent strain. The same clinical study finds the ATE among younger people to be 10%, as 55% survive when vaccinated and 45% survive when unvaccinated. If just comparing ATEs, it would seem that the vaccine is $20\% - 10\% = 10$ percentage points more effective among older people. In this case we would be comparing effects of the vaccine cause, an EoC analysis.

The quantity of interest we really care about is whether the vaccine is the cause of survival, a CoE analysis. This quantity is $P(\text{benefit})$. We can then compare $P(\text{benefit})$ for elders versus young. Traditional counterfactual analysis [TP00] yields bounds on $P(\text{benefit})$ among the sixty-and-over group of 20% to 57%. This is a large range and it starts to become clear that our true quantity of interest is not necessarily what we would think with an EoC analysis. Bounds on $P(\text{benefit})$ among the younger group is calculated to be between 10% and 55%. This is also a large range and it significantly overlaps with the $P(\text{benefit})$ bounds among the older subpopulation. Which group should receive priority for vaccination?

As discussed in Chapter 3, we have an additional tool in our CoE arsenal, the ability to use observational data in addition to experimental data. Remarkably, taking into account

individuals' whims and desires, such as their willingness to get vaccinated, through observational data, can narrow bounds on probabilities of causation. In some cases, this narrowing can be so acute that it leads to point estimates. As will be seen in Section 4.2, realistic observational data can result in bounds of $[20\%, 40\%]$ for over-fifty-niners and $[40\%, 55\%]$ for under-sixtiers. This would reverse our naïve vaccine prioritization under the EoC analysis above.

While the above demonstrates value in existing methods to compute bounds on $P(\text{benefit})$, often these existing methods cannot sufficiently narrow the bounds enough to improve policies or decisions. However, additional population-level data on covariates and mild structural assumptions on the causal graph can further narrow those bounds significantly.

This chapter presents methods to compute narrower bounds on the PoCs mentioned in Chapter 3 as well as another two PoCs: Probability of Necessity (PN) and Probability of Sufficiency (PS). Beyond causal effects, it is surprising that the structure of the causal graph allows us to narrow these bounds. The graph describes properties of the population, yet adds information about individuals. In this way, individual level effects are obtained from population data. Section 4.3 will demonstrate this with covariates even when they are not needed for identification of causal effects.

The next sections are organized as follows. Section 4.2 offers descriptions and existing analyses of the six PoCs. Section 4.3 covers narrowing these PoCs using covariate data, including formulas, proofs, and graphical criterion. Section 4.4 provides the same type of analyses for mediators. Section 4.5 gives tools to combine multiple covariate and mediator data with more complicated graphs to further narrow bounds. Finally, Section 4.6 concludes with a discussion of the results.

4.2 Probability of Necessity and Probability of Sufficiency

Two important PoCs, as defined in [Pea99], are PN and PS. Causal diagrams [PEA95, SGS01, Pea09, KF09] and the language of counterfactuals in its structural model semantics, as given in [BP95, GP98, Hal00] are used in the following sections.

4.2.1 Probability of Necessity

Assume you went to the beach and acquired Covid-19. Was it necessarily the exposure you had at the beach which caused you to acquire the disease? The probability that you would not have acquired Covid-19 had you not gone to the beach, given that you did in fact go to the beach and acquired it, is called the Probability of Necessity (PN). This clearly has important implications for public health policy, as for risk assessment and reflection on a personal level.

Definition 4.2.1 (Probability of Necessity (PN))**.** *PN is defined [Pea99] as the expression:*

$$\begin{aligned} PN &\triangleq P(Y_{x'} = false | X = true, Y = true) \\ &= P(y'_{x'} | x, y). \end{aligned} \tag{4.1}$$

In other words, PN stands for the probability that event y would not have occurred in the absence of event x, given that x and y did in fact occur.

PN has applications in epidemiology, legal reasoning, and artificial intelligence. Epidemiologists have long been concerned with estimating the probability that a certain case of disease is attributable to a particular exposure, which is normally interpreted counterfactually as "the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur." This counterfactual notion is also used frequently in lawsuits, where legal responsibility is at the center of contention.

4.2.2 Probability of Sufficiency (PN)

Contrary to the scenario above with PN, assume you stayed home and didn't acquire Covid-19. Would going to the beach have been sufficient to acquire the disease? The probability that you would have acquired Covid-19 had you gone to the beach, given that you stayed home and did not acquire it, is called the Probability of Sufficiency (PS). This is essentially the converse of PN.

Definition 4.2.2 (Probability of Sufficiency (PS))**.** *[Pea99]*

$$PS \triangleq P(y_x|y', x'). \tag{4.2}$$

PS finds applications in policy analysis, artificial intelligence, and psychology. A policy maker may well be interested in the dangers that a certain exposure may present to the healthy population [KFG89]. Counterfactually, this notion is expressed as the "probability that a healthy unexposed individual would have gotten the disease had he/she been exposed." In psychology, PS serves as the basis for Cheng's [Che97] causal power theory [Gly13], which attempts to explain how humans judge causal strength among events. In artificial intelligence, PS plays a major role in the generation of explanations [Pea09].

4.2.3 Bounds

The two probabilities mentioned above, PN and PS, as well as the four PoCs mentioned previously, P(benefit), P(harm), P(immunity), and P(doom), are counterfactual notions, for they pertain to the behavior of an individual patient under two incompatible conditions. As such, they usually can't be estimated precisely from group data, even when experimental and observational data are available for all variables, regardless of how big the data is. As was the case the the four previous PoCs, informative bounds for PN and PS can be derived when experimental and observational data are available. However, informative bounds are not available for PN and PS if only experimental or only observational data are available,

unlike the four previous PoCs. These bounds were produced and proven tight, in the sense of being the narrowest possible given the data, by Tian and Pearl [TP00, Pea09] through a linear program. Li and Pearl [LP19] provide a theoretical proof of the tight bounds for PN, PS, and P(benefit) (equivalent to P(benefit)), and other probabilities of causation without a causal diagram.

The following bounds will be referred to as Tian-Pearl bounds for the remainder of this dissertation:

$$\max\left\{0, \frac{P(y) - P(y_{x'})}{P(x,y)}\right\} \leqslant \text{PN} \leqslant \min\left\{1, \frac{P(y'_{x'}) - P(x',y')}{P(x,y)}\right\}, \tag{4.3}$$

$$\max\left\{0, \frac{P(y_x) - P(y)}{P(x',y')}\right\} \leqslant \text{PS} \leqslant \min\left\{1, \frac{P(y_x) - P(x,y)}{P(x',y')}\right\}. \tag{4.4}$$

Bounds for a specific subpopulation, defined by a set C of pretreatment characteristics, can be obtained by simply conditioning each probability in the bounds above on $C = c$.

4.2.3.1 Exogeneity

Definition 4.2.3 (Exogeneity)**.** *A variable X is said to be exogenous for Y in model M iff [TP00]*

$$Y_x \perp\!\!\!\perp X \quad \textit{and} \quad Y_{x'} \perp\!\!\!\perp X. \tag{4.5}$$

In other words, the way Y would potentially respond to experimental conditions x or x' is independent of the actual value of X.

If exogeneity holds, then the bounds on PN, PS, and P(benefit) become:

$$\frac{\max\{0, P(y|x) - P(y|x')\}}{P(y|x)} \leqslant \text{PN} \leqslant \frac{\min\{P(y|x), P(y'|x')\}}{P(y|x)}, \tag{4.6}$$

$$\frac{\max\{0, P(y|x) - P(y|x')\}}{P(y'|x')} \leqslant \text{PS} \leqslant \frac{\min\{P(y|x), P(y'|x')\}}{P(y'|x')}, \tag{4.7}$$

$$\max\{0, P(y|x) - P(y|x')\} \leqslant P(\text{benefit}) \leqslant \min\{P(y|x), P(y'|x')\}. \tag{4.8}$$

Remarkably, Tian and Pearl [TP00] proved that strong exogeneity, $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, does not improve the bounds.

Because P(harm) is related to P(benefit) through the ATE, which is $P(y|x) - P(y|x')$ under exogeneity, and P(immunity) and P(doom) are really to P(benefit) through causal effects $P(y_x) = P(y|x)$ and $P(y'_x) = P(y'|x)$ (under exogeneity), the remaining three PoCs are easily derivable under exogeneity:

$$\max\{0, P(y|x') - P(y|x)\} \leqslant P(\text{harm}) \leqslant \min\{P(y'|x), P(y|x')\}, \tag{4.9}$$

$$\max\{0, P(y|x) - P(y'|x')\} \leqslant P(\text{immunity}) \leqslant \min\{P(y|x), P(y|x')\}, \tag{4.10}$$

$$\max\{0, P(y'|x) - P(y|x')\} \leqslant P(\text{doom}) \leqslant \min\{P(y'|x), P(y'|x')\}. \tag{4.11}$$

Conceptually, causal models should be informative for P(benefit), P(harm), P(immunity), and P(doom) whenever they enable the identification of causal effects, for then they allow us to substitute those effects for experimental data that are needed for computing the Tian-Pearl bounds. This substitution, however, is only one way in which causal models can be leveraged to assess these PoCs.

Sections 4.3, 4.4, and 4.5 will go a step further and derive even *narrower* bounds when structural information is available in the form of a causal model, or properties of such a model. Model-based information was used in estimating the extent to which radiation was responsible for leukemia [Pea09, pages 299-301] and, more recently, for attributing individual responsibility in legal settings [DMM17].

4.2.4 Benefit Example

Imagine a randomized controlled trial (RCT) for a new treatment of a disease conducted with a treatment group and a non-treatment (control) group for one week. Among the treated, 40% are cured and 60% remain sick or die. Exactly the same proportions are found in the control group (those who were denied treatment), 40% are cured, and 60% remain sick or die. This treatment would be deemed ineffective by the FDA and other public policy makers. If you had the disease, you would probably be reluctant to undertake this treatment, especially if there are significant costs or side-effects. However, given the severity of your

condition, your family history, and other idiosyncratic dispositions you may be inclined to try it anyhow. Isn't it possible, you might hope, that this treatment actually *cures* 40% of patients in a category similar to yours and kills 40% of patients which are dissimilar to you. What you really want to know, and an RCT usually can't tell us, is the probability that you are in need of such treatment, i.e., cured if treated and and not cured when not treated.

Is it possible that the treatment actually cured 40% of patients? The ATE is zero, but the Tian-Pearl bounds can be calculated to see what the possible $P(\text{benefit})$ probabilities are. Observational data are unavailable, so only the first two arguments to min and max of Equations (3.6) will be used:

$$\begin{aligned}
\max\{0, P(y_x) - P(y_{x'})\} \leqslant P(\text{benefit}) &\leqslant \min\{P(y_x), P(y'_{x'})\} \\
\max\{0, 0.4 - 0.4\} \leqslant P(\text{benefit}) &\leqslant \min\{0.4, 0.6)\} \\
0 \leqslant P(\text{benefit}) &\leqslant 0.4.
\end{aligned}$$

Therefore, the probability that an individual would be cured with treatment and not cured without treatment is between 0 and 0.4. Since this includes 40%, it is possible the treatment cured 40% of patients. That's a significant portion of the population, while a naïve interpretation of the RCT results would have deemed the treatment ineffective. A lesson from *Causal Inference: What If* [HR20, page 6] is, "Absence of an average causal effect does not imply absence of individual effects."

If the treatment does cure 40% of patients, the reason this 40% is not reflected in the ATE is that the treatment must have caused 40% of patients to not be cured. Had they not been treated they would have naturally been cured. The beneficial effects of the treatment were canceled out by the harmful effects. The ramifications on personal decision making are serious [MP21] and the RCT hid this information. The treatment is not so safe with up to 40% of patients being harmed by it. Physicians and policy makers would need to be aware of this. With the potential to save 40% of patients' lives, further research certainly warrants attention and investment to distinguish subpopulations benefiting from subpopulations being

harmed by treatment.

4.2.5 Identification

Monotonicity is the condition that treatment never harms patients:

Definition 4.2.4 (Monotonicity)**.** *A variable Y is said to be monotonic relative to variable X in a causal model M iff [TP00]*

$$y'_x \wedge y_{x'} = \mathit{false}. \tag{4.12}$$

In the RCT above, monotonicity is satisfied when it is impossible for an individual to be not cured if given treatment and cured if not given treatment. In other words, when $P(\text{harm}) = 0$. In this case, point estimates, rather than bounds, can be found for all six PoCs discussed so far:

$$\text{PN} = \frac{P(y) - P(y_{x'})}{P(x, y)}, \tag{4.13}$$

$$\text{PS} = \frac{P(y_x) - P(y)}{P(x', y')}, \tag{4.14}$$

$$P(\text{benefit}) = P(y_x) - P(y_{x'}), \tag{4.15}$$

$$P(\text{harm}) = P(y_{x'}) - P(y_x),$$

$$P(\text{immunity}) = P(y_x) - P(y'_{x'}),$$

$$P(\text{doom}) = P(y'_x) - P(y_{x'}).$$

The intuition behind this is best exemplified in the $P(\text{benefit})$ estimate. Individuals in the treatment group with a positive outcome could be benefiters or immune. Which type is unknown because of the inability to peek at how those individuals would have behaved had they not been treated. Therefore, $P(y_x)$ is the proportion of individuals belonging to the

benefit or immunity groups. Similarly, $P(y'_x)$ is the proportion of individuals belonging to the harm or immunity groups.

The ATE, defined as $P(y_x) - P(y_{x'})$, is the proportion of benefiters and immune minus the proportion of harmed and immune, leaving us with the proportion of benefiters minus the proportion of harmed. Therefore, $P(\text{benefit}) = P(y_x) - P(y_{x'}) + P(\text{harm})$. This is why the ATE was only a lower bound for $P(\text{benefit})$. If monotonicity is assumed, then we finally arrive at the point estimate $P(\text{benefit}) = P(y_x) - P(y_{x'})$.

Note that in the Tian-Pearl bounds, bounds under exogeneity, and identification, PN and PS are swapped by simply exchanging x with x' and y with y' due to their converse nature. Since equations and algorithms for PN can so easily be converted to be for PS, and similarly, the equations and algorithms for $P(\text{benefit})$ can be converted to be for $P(\text{harm})$, $P(\text{immunity})$, and $P(\text{doom})$, the remainder of this chapter will only consider PN and $P(\text{benefit})$ and consider PS only sparingly.

4.3 Leveraging Covariate Data

The techniques profiting from covariate data will commence with identification of PN and $P(\text{benefit})$ under monotonicity, as defined in Definition 4.2.4. This will be followed by an exploration of formulas to compute narrower bounds with an admissible covariate set [PP10]. A set of covariates is admissible when they satisfy the back-door criterion. Without an admissible covariate set, bounds can still be narrowed with an inadmissible covariate set, especially if both experimental and observational data are available.

4.3.1 Observational Data Under Monotonicity

Tian and Pearl [TP00] pointed out that the PN, PS, and $P(\text{benefit})$ are identifiable, under the monotonicity assumption, if the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable. Those causal effects could be directly obtained through experimental data or they could be identified

through adjustment from covariate data. For example, if covariate set $\mathbf{Z}$ satisfies the back-door criterion [Pea93], then we can identify $P(y_x)$ and $P(y_{x'})$ and, therefore, PN, PS, and P(benefit) under monotonicity:

$$P(y_x) = \sum_{\mathbf{z}} P(y|x,\mathbf{z}) \cdot P(\mathbf{z}),$$
$$P(y_{x'}) = \sum_{\mathbf{z}} P(y|x',\mathbf{z}) \cdot P(\mathbf{z}).$$

PN only requires $P(y_{x'})$ to be identified, while PS only requires $P(y_x)$ to be identified:

$$\text{PN} = \frac{P(y) - \sum_{\mathbf{z}} P(y|x',\mathbf{z}) \cdot P(\mathbf{z})}{P(x,y)},$$
$$\text{PS} = \frac{\sum_{\mathbf{z}} P(y|x,\mathbf{z}) \cdot P(\mathbf{z}) - P(y)}{P(x',y')},$$
$$P(\text{benefit}) = \mathbb{E}_{\mathbf{z}}[P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})]. \tag{4.16}$$

Kuroki and Cai [KC11] simplified these equations for PN and P(benefit) by first defining a PN and P(benefit) stratified by a set of covariates $\mathbf{Z}$:

$$\text{PN}(\mathbf{z}) = P(y'_{x'}|x,y,\mathbf{z}), \tag{4.17}$$
$$P(\text{benefit}|\mathbf{z}) = P(y_x, y'_{x'}|\mathbf{z}). \tag{4.18}$$

Under monotonicity, these become:

$$\text{PN}(\mathbf{z}) = \frac{P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z})}{P(x,y|\mathbf{z})},$$
$$P(\text{benefit}|\mathbf{z}) = P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}).$$

Let $\text{PN}_{\mathbf{z}}$ and $P(\text{benefit})_{\mathbf{z}}$ be the PN and P(benefit), respectively, when evaluating with the set of covariates $\mathbf{Z}$:

$$\text{PN}_{\mathbf{z}} = \sum_{\mathbf{z}} \text{PN}(\mathbf{z}) \cdot P(\mathbf{z}|x,y), \tag{4.19}$$
$$P(\text{benefit}|\mathbf{z}) = \sum_{\mathbf{z}} P(\text{benefit}|\mathbf{z}) \cdot P(\mathbf{z})). \tag{4.20}$$

When $\mathbf{Z}$ satisfies the back-door criterion, PN, PS (through an easy derivation from PN), and P(benefit) require only observational data:

$$\text{PN}_{\mathbf{z}} = \mathbb{E}_{\mathbf{z}}[P(y|\mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(x, y)^{-1}, \tag{4.21}$$

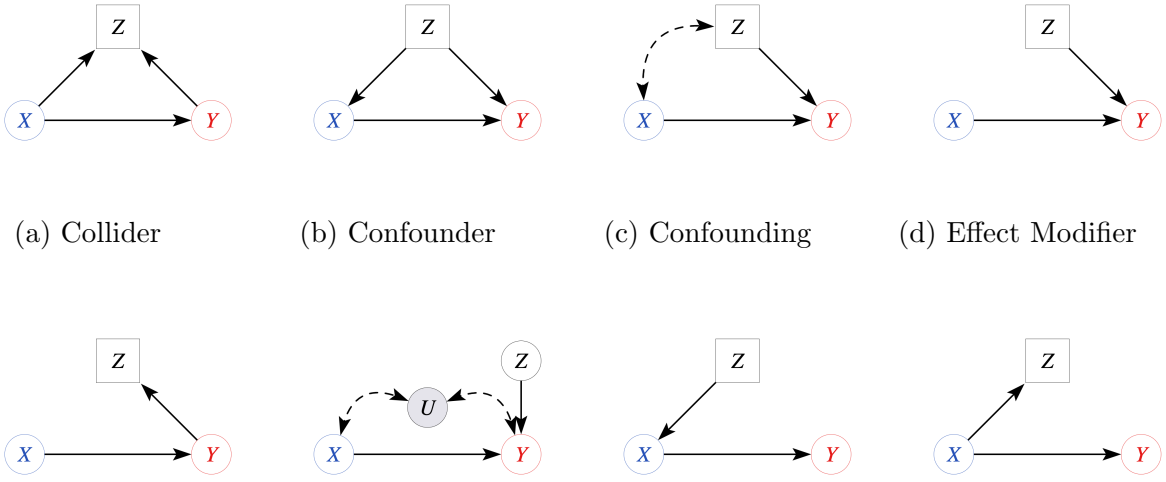
$$P(\text{benefit})_{\mathbf{z}} = \mathbb{E}_{\mathbf{z}}[P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})]. \tag{4.22}$$

In this way, covariate data has enabled us to identify PN, PS, and P(benefit) in the absence of experimental data. However, monotonicity is a strong and necessary assumption. Kuroki and Cai's stratified PN and P(benefit) will next be used to relax this monotonicity assumption.

4.3.2 Admissible Covariates

Following the observations of settings [DMM17], the role of causal models will be shown to extend beyond identification; they may actually enable us to narrow the PN, PS, and P(benefit) bounds even in situations where identification is neither feasible nor needed, such as when experimental data are available. The purpose of this chapter is to understand the role that causal models can play in the transition from group data to individual behavior and, more concretely, to define the conditions under which measurements of covariates in the model may narrow the bounds for PN, PS, and P(benefit). A typical covariate, in the context of the beach-going RCT study of sections 4.2.1 and 4.2.2, would be pre-treatment variables in both treatment and control groups and asking whether it provides a more accurate assessment of PN, PS, and P(benefit) for a typical individual not in the study. Section 4.4 will examine the same question for post-treatment side effects.

The following analysis is based on the bounds derived in [TP00] and parallels and extends the analyses of [MLP22, DMM17] for the models described in Figure 4.1. More complex models can be constructed from the graphical criterion presented below.



(a) Collider (b) Confounder (c) Confounding (d) Effect Modifier

(e) Effect of Effect (f) M-bias (g) Pre-treatment (h) Treatment Effect

Figure 4.1: Core conditional ignorability DAG structures

Exogeneity holds in each stratum $\mathbf{z}$ of $\mathbf{Z}$ iff

$$P(y_x|\mathbf{z}) = P(y|x,\mathbf{z}) \quad \text{and} \quad P(y_{x'}|\mathbf{z}) = P(y|x',\mathbf{z}),$$

in other words, when conditional ignorability holds. Conditional ignorability imposes a demand on a causal structure in order to measure $P(y_x|\mathbf{z})$ and $P(y_{x'}|\mathbf{z})$. In particular, $\mathbf{Z}$ cannot contain any descendants of X, unless $Y_x \perp\!\!\!\perp \mathbf{Z}_{X\text{-descendant}}$, where $\mathbf{Z}_{X\text{-descendant}}$ is the subset of Z consisting of descendants of X. The reason for this constraint is if X was set to x and $\mathbf{Z}$ contains a descendant of X, then $\mathbf{Z}$ could be altered. Then $P(y_x|\mathbf{z})$ and $P(y_{x'}|\mathbf{z})$ would likely be unmeasurable counterfactual terms. A more explicit probability equivalent to $P(y_x|\mathbf{z})$ is $P(y_{x,\mathbf{Z}_x}|\mathbf{z})$, where $\mathbf{Z}_x$ makes $P(y_{x,\mathbf{Z}_x}|\mathbf{z})$ the probability of $Y = y$ had X been set to x and $\mathbf{Z}$ was set to its natural value after setting X to x, conditioned on $\mathbf{Z} = \mathbf{z}$. One scenario that allows a causal Bayesian network to infer $P(y_x|\mathbf{z})$ when $\mathbf{Z}$ contains descendants of X are when the conditional probability tables (CPTs) for $\mathbf{Z}$ and their ancestors are deterministic (probabilities are 0 or 1). If all endogenous variables

are deterministic and marginal probabilities of exogenous variables are known, then the original counterfactual probability is estimable. Another scenario where $P(y_x|\mathbf{z})$ is estimable is when $P(\mathbf{z}_x) = 1$. If the descendants in $\mathbf{Z}$ are independent of Y_x, then $P(y_x|\mathbf{z})$ would be measurable, but those descendants would not contribute to any narrowing of bounds. In the rare circumstance that the functional model or SCM is available, counterfactual terms can be computed. However, an SCM also allows direct computation of the PN, PS, and P(benefit), so the bounds described below are unnecessary. Descendants of X are allowed in the covariate set if $Y_x \perp\!\!\!\perp \mathbf{Z}_{X\text{-descendant}}$, as in Figure 4.1h, because, coupled with conditional ignorability, this implies $P(y_x|\mathbf{z}) = P(y|x, \mathbf{z})$ and $P(y_{x'}|\mathbf{z}) = P(y|x', \mathbf{z})$, which are measurable.

However, it will be clear below that there is no bound-narrowing advantage in including $\mathbf{Z}_{X\text{-descendant}}$ among the set of covariates. In fact, Cinelli, Forney, and Pearl [CFP24, page 7] point out that conditioning on Z of Figure 4.1h reduces variation in X. This can hurt ATE precision in finite samples.

In the case of figures 4.1d, 4.1e, and 4.1f it seems that measurements of $\mathbf{Z}$ are superfluous, since they are not needed for deconfounding X and Y. However, it will be shown that such measurement may nevertheless improve the bounds (4.6), (4.7), and (4.8).

Sections 4.3.2.1 and 4.3.2.2 assume $\mathbf{Z}$ satisfies the back-door criterion, rendering X exogenous for Y in each stratum $\mathbf{z}$ of $\mathbf{Z}$.

4.3.2.1 PN Bounds

With the assumption that conditioning on $\mathbf{Z}$ leaves X exogenous, the bounds (4.6) can be applied to PN($\mathbf{z}$) in order to obtain:

$$\max\left\{0, 1 - \frac{P(y|x', \mathbf{z})}{P(y|x, \mathbf{z})}\right\} \leqslant \text{PN}(\mathbf{z}) \leqslant \min\left\{1, \frac{P(y'|x', \mathbf{z})}{P(y|x, \mathbf{z})}\right\}. \tag{4.23}$$

The task is now to bound $\text{PN}_{\mathbf{z}}$, the population PN evaluated with covariate set $\mathbf{Z}$, using the bounds derived at (4.23) for the $\mathbf{z}$-specific PN($\mathbf{z}$). By replacing PN($\mathbf{z}$) in the summation

within Equation (4.19) by its lower bound in (4.23), the lower bound for PN is as follows:

$$\begin{aligned}\mathrm{PN}_{\mathbf{z}} &\geqslant \sum_{\mathbf{z}} \mathrm{PN}_{\text{lower-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\ &= P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \end{aligned} \tag{4.24}$$

Similarly, $\mathrm{PN}(\mathbf{z})$ in the summation within Equation (4.19) can be replaced by its upper bound in (4.23) to obtain an upper bound as follows:

$$\begin{aligned}\mathrm{PN}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \mathrm{PN}_{\text{upper-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\ &= 1 - P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y'|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \end{aligned} \tag{4.25}$$

Note that $P(\mathbf{z}|x)$ can be simplified to $P(\mathbf{z})$ in (4.24) and (4.25) if $\mathbf{Z} \perp\!\!\!\perp X$, as in figures 4.1d and 4.1f.

Bounds taking advantage of $\mathbf{Z}$ will always be within Tian-Pearl bounds. In parts, the lower bound of $\mathrm{PN}_{\mathbf{z}}$ will be greater or equal to the Tian-Pearl lower bound and the upper bound of $\mathrm{PN}_{\mathbf{z}}$ will be less than or equal to the Tian-Pearl upper bound. The following lemma is necessary to show this superiority:

Lemma 4.3.1. *Given two n-length sequences, $\langle a_1, a_2, \ldots, a_n\rangle$ and $\langle b_1, b_2, \ldots, b_n\rangle$, the maximum between the summation of $\langle a_1, a_2, \ldots, a_n\rangle$ and the summation of $\langle b_1, b_2, \ldots, b_n\rangle$ will always be less than or equal to the summation of the maximum between each a_i and b_i, where i is the index in the sequence:*

$$\max\left\{\sum_i a_i, \sum_i b_i\right\} \leqslant \sum_i \max\left\{a_i, b_i\right\}. \tag{4.26}$$

Similarly, the minimum between the summation of $\langle a_1, a_2, \ldots, a_n\rangle$ and the summation of $\langle b_1, b_2, \ldots, b_n\rangle$ will always be greater than or equal to the summation of the minimum between each a_i and b_i:

$$\min\left\{\sum_i a_i, \sum_i b_i\right\} \geqslant \sum_i \min\left\{a_i, b_i\right\}. \tag{4.27}$$

Both (4.26) *and* (4.27) *will be equality when $\forall i : a_i \leqslant b_i$ or $\forall i : a_i \geqslant b_i$.*

Let us compare the Tian-Pearl PN lower bound of (4.6) with $\text{PN}_{\mathbf{z}}$'s lower bound:

$$\begin{aligned}
\text{PN} &\geqslant \max\left\{0, 1 - \sum_{\mathbf{z}} \frac{P(y|x', \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y)\right\} \\
&= \max\left\{0, \sum_{\mathbf{z}} \frac{P(y|x, \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y) - \sum_{\mathbf{z}} \frac{P(y|x', \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y)\right\} \\
&= \max\left\{0, \sum_{\mathbf{z}} \frac{P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})}{P(y|x)} P(\mathbf{z}|x)\right\} \\
&= P(y|x)^{-1} \max\left\{0, \sum_{\mathbf{z}} [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}|x)\right\}, \qquad (4.28) \\
\text{PN}_{\mathbf{z}} &\geqslant P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}|x)\right\} \qquad (4.29)
\end{aligned}$$

The inequality (4.26), with $a_i = 0$ and $b_i = [P(y|x, \mathbf{z}_i) - P(y|x', \mathbf{z}_i)] \cdot P(\mathbf{z}_i|x)$, shows the Tian-Pearl lower bound in (4.28) is inferior to $\text{PN}_{\mathbf{z}}$'s lower bound in (4.29). Let us now compare the Tian-Pearl PN upper bound in (4.6) with $\text{PN}_{\mathbf{z}}$'s upper bound in (4.25):

$$\begin{aligned}
\text{PN} &\leqslant \min\left\{1, \sum_{\mathbf{z}} \frac{P(y'|x', \mathbf{z})}{P(y|x, \mathbf{z})} \cdot P(\mathbf{z}|x, y)\right\} \\
&= \min\left\{\sum_{\mathbf{z}} P(\mathbf{z}|x), \sum_{\mathbf{z}} \frac{P(y'|x', \mathbf{z})}{P(y|x)} \cdot P(\mathbf{z}|x)\right\}, \qquad (4.30) \\
\text{PN}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x', \mathbf{z})}{P(y|x)}\right\} \cdot P(\mathbf{z}|x) \\
&= \sum_{\mathbf{z}} \min\left\{P(\mathbf{z}|x), \frac{P(y'|x', \mathbf{z})}{P(y|x)} \cdot P(\mathbf{z}|x)\right\}. \qquad (4.31)
\end{aligned}$$

The inequality (4.27), with $a_i = P(\mathbf{z}_i|x)$ and $b_i = \frac{P(y'|x', \mathbf{z}_i)}{P(y|x)} \cdot P(\mathbf{z}_i|x)$, shows the Tian-Pearl upper bound in (4.30) is inferior to $\text{PN}_{\mathbf{z}}$'s lower bound in (4.31).

From lemma 4.3.1, there is no bounds narrowing advantage using covariate set $\mathbf{Z}$ when $\forall i : a_i \leqslant b_i$ or $\forall i : a_i \geqslant b_i$. For the lower bound of $\text{PN}_{\mathbf{z}}$ this means, $\forall i$:

$$\begin{aligned}
0 &\leqslant [P(y|x, \mathbf{z}_i) - P(y|x', \mathbf{z}_i)] \cdot P(\mathbf{z}_i|x), \\
P(y|x', \mathbf{z}_i) &\leqslant P(y|x, \mathbf{z}_i),
\end{aligned}$$

or $\forall i : P(y|x', \mathbf{z}_i) \geqslant P(y|x, \mathbf{z}_i)$.

There is no smaller upper bound advantage using covariate set $\mathbf{Z}$ when, $\forall i$:

$$P(\mathbf{z}_i|x) \leqslant \frac{P(y'|x', \mathbf{z}_i)}{P(y|x)} \cdot P(\mathbf{z}_i|x),$$
$$P(y|x) \leqslant P(y'|x', \mathbf{z}_i),$$

or $\forall i : P(y|x) \geqslant P(y'|x', \mathbf{z}_i)$.

With the assumption that $\mathbf{Z}$ satisfies the back-door criterion, bounds on PN can be narrowed from observational data on X, Y, and Z. Kuroki and Cai [KC11] extended the monotonicity assumption to conditional monotonicity, expressed as $P(y_{x'}, y'_x|z) = 0$. In stratum where conditional monotonicity holds, both the lower and upper PN bound can improve further by using Equation 4.13 instead of min or max.

Attention now turns to the next probability of causation with back-door criterion satisfying covariates. Then graphical criterion in Section 4.3.2.3 will graphically demonstrate when these bounds can be used. Finally, examples will illustrate the application of covariate-benefiting bounds.

4.3.2.2 P(**benefit**) **Bounds**

Tian and Pearl [TP00] provided bounds on P(benefit) for observational-only data with no assumptions of exogeneity. While there is no effective lower bound (the lower bound remains 0, as with all probabilities), an upper bound can be informative:

$$0 \leqslant P(\text{benefit}) \leqslant P(x, y) + P(x', y').$$

Narrower bounds than this cannot be obtained using the summation of maximums or minimums technique in lemma 4.3.1 because there different minimum or maximum options to choose from do not exist in each stratum of Z.

With the same conditional ignorability assumption of Section 4.3.2.1, the bounds in

Equation (4.8) can be applied to $P(\text{benefit}|\mathbf{z})$ in order to obtain:

$$\max\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\} \leqslant P(\text{benefit}|\mathbf{z}) \leqslant \min\{P(y|x,\mathbf{z}), P(y'|x',\mathbf{z})\}. \tag{4.32}$$

With the same approach taken in Section 4.3.2.1, the $P(\text{benefit})_{\mathbf{z}}$ lower bound is as follows:

$$\begin{aligned} P(\text{benefit})_{\mathbf{z}} &\geqslant \sum_{\mathbf{z}} P(\text{benefit}|\mathbf{z})_{\text{lower-bound}} \cdot P(\mathbf{z}) \\ &= \sum_{\mathbf{z}} \max\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\} \cdot P(\mathbf{z}) \qquad (4.33) \\ &= \sum_{\mathbf{z}} \max\{0, [P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(\mathbf{z})\}. \qquad (4.34) \end{aligned}$$

The $P(\text{benefit})_{\mathbf{z}}$ upper bound is analogously:

$$\begin{aligned} P(\text{benefit})_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} P(\text{benefit}|\mathbf{z})_{\text{upper-bound}} \cdot P(\mathbf{z}) \\ &= \sum_{\mathbf{z}} \min\{P(y|x,\mathbf{z}), P(y'|x',\mathbf{z})\} \cdot P(\mathbf{z}) \qquad (4.35) \\ &= \sum_{\mathbf{z}} \min\{P(y|x,\mathbf{z}) \cdot P(\mathbf{z}), P(y'|x',\mathbf{z}) \cdot P(\mathbf{z})\}. \qquad (4.36) \end{aligned}$$

In comparison with Tian-Pearl bounds:

$$\begin{aligned} P(\text{benefit}) &\geqslant \max\{0, \sum_{\mathbf{z}} [P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(\mathbf{z})\}, \\ P(\text{benefit}) &\leqslant \min\{P(y|x,\mathbf{z}) \cdot P(\mathbf{z}), P(y'|x',\mathbf{z}) \cdot P(\mathbf{z})\}. \end{aligned}$$

Lemma 4.3.1 makes the superiority of the new bounds clear. The lower bound will be higher than the Tian-Pearl $P(\text{benefit})$ lower bound when $P(y_x|Z=z) - P(y_{x'}|Z=z)$ is greater than 0 for some Z and less than 0 for other Z. Similarly, the upper bound will be lower than the Tian-Pearl $P(\text{benefit})$ upper bound when $P(y_x|Z=z) > P(y'_{x'}|Z=z)$ for some Z and $P(y_x|Z=z) < P(y'_{x'}|Z=z)$ for other Z.

4.3.2.3 Graphical Criterion

The only assumptions made in this section for bounds on $\text{PN}_{\mathbf{z}}$ and $P(\text{benefit})_{\mathbf{z}}$ are:

- $\mathbf{Z}$ satisfies the back-door criterion relative to (X, Y)
- No node in $\mathbf{Z}$ is a descendant of X, unless they are independent of Y_x

In Section 4.3.3 the first assumption will be relaxed. Additionally, no advantage can be expected on narrowing bounds if $Y \perp\!\!\!\perp Z \mid X$, as in figures 4.1g and 4.1h. This is because all of the probabilities inside the max and min functions of the form $P(y|x, z)$ become $P(y|x)$, reducing bounds on $\text{PN}_{\mathbf{z}}$ and $P(\text{benefit})_{\mathbf{z}}$ to the Tian-Pearl bounds on PN and $P(\text{benefit})$.

4.3.2.4 PN Example

A new pharmaceutical drug purportedly has a side effect of debilitating pain for months. A particular person takes the drug and, unfortunately, experiences outrageous pain that affects their job, family, and sanity. What is it necessarily the drug that caused this person to suffer so much?

This is a PN query. Let Y represent this horrible pain with y meaning the pain was experienced and y' meaning the pain was not experienced. Let X represent the drug with x meaning the drug was taken and x' meaning the drug was not taken. For simplicity, this example will use a single binary covariate. Let Z represent a medical condition with z meaning the condition is present and z' meaning the condition is absent. Researchers are confident that Z satisfies conditional ignorability. The graph associated with this scenario is depicted in Figure 4.1b. This means Z can be used as a covariate to more narrowly bound PN through observational data alone.

Among people highly susceptible to these excruciating and lengthy bouts of pain, it turns out that this medical condition Z acts as a protective agent: 20% will endure the severe suffering if they have the medical condition versus 80% without the condition. Doctors

observe that people with this medical condition who take the new pharmaceutical drug experience debilitating pain 60% of the time, so it seems the drug might remove some of the protective mechanism. Of those without the medical condition, only 40% of drug-takers endure the pain. The same number of people who take drug have the medical condition and do not have the medical condition. These proportions are reflected in the conditional probabilities of Table 4.1.

	Conditioned on x	Conditioned on x'
$P(z)$	0.5	unknown
$P(y\|z)$	0.6	0.2
$P(y\|z')$	0.4	0.8

Table 4.1: Conditional probabilities for PN example

First, the Tian-Pearl bounds will be calculated from this data. Then the narrower bounds that take advantage of conditioning on Z will be calculated. Tian-Pearl bounds yield,

$$\max\left\{0, \frac{0.5-0.5}{0.5}\right\} \leqslant \text{PN} \leqslant \frac{\min\{0.5, 0.5\}}{0.5},$$
$$0 \leqslant \text{PN} \leqslant 1.$$

Thus, Tian-Pearl bounds provide no information for PN. Utilizing measurements of Z, on the other hand, gives the lower bound from (4.24):

$$\begin{aligned}\text{PN}_{\mathbf{z}} &\geqslant P(y|x)^{-1} \cdot \sum_z \max\{0, P(y|x,z) - P(y|x',z)\} \cdot P(z|x) \\ &= 0.5^{-1} \cdot (\max\{0, 0.6-0.2\} \cdot 0.5 + \max\{0, 0.4-0.8\} \cdot 0.5) \\ &= 0.4.\end{aligned}$$

The upper bound from (4.25) is:

$$\begin{aligned}\mathrm{PN}_{\mathbf{z}} &\leqslant 1 - P(y|x)^{-1} \cdot \sum_{z} \max\{0, P(y|x,z) - P(y'|x',z)\} \cdot P(z|x) \\ &= 1 - 0.5^{-1} \cdot (\max\{0, 0.6 - 0.8\} \cdot 0.5 + \max\{0, 0.4 - 0.2\} \cdot 0.5) \\ &= 0.8.\end{aligned}$$

Thus, the new bounds are $0.4 \leqslant \mathrm{PN}_{\mathbf{z}} \leqslant 0.8$, thanks to measurement of Z. This is tremendously more informative than the Tian-Pearl bounds. More extreme examples can demonstrate a range reduction of 1 ($0 \leqslant PN \leqslant 1$) to 0, namely, a precise value for PN. Clearly the bounds narrowing can be significant.

A person might be making a decision of whether to blame the pharmaceutical company for their long-lasting debilitating pain. Blame might mean suing the company or, at least, publicly shaming the company. But first, they need to know the true probability that the drug was a necessary cause. Traditional PN bounds were uninformative. The $\mathrm{PN}_{\mathbf{z}}$ bounds made the likelihood reasonable enough to pursue the company.

Note that the person doesn't know their medical condition Z. This is a crucial point. Had they known their medical condition status, they would just use that data. This matter will be revisited in Section 4.3.5.

4.3.3 Combined Data

One of the remarkable results of Tian-Pearl bounds with combined observational and experimental data is that it is not only valid to have confounding, but confounding can actually narrow the bounds. This section will mirror the analysis in Section 4.3.2 with the conditional ignorability requirement waived. A non-null set of covariates will still need to be used, but additional unobserved confounding can exist. An example of remaining confounding after conditioning on a covariate is displayed in Figure 4.2.

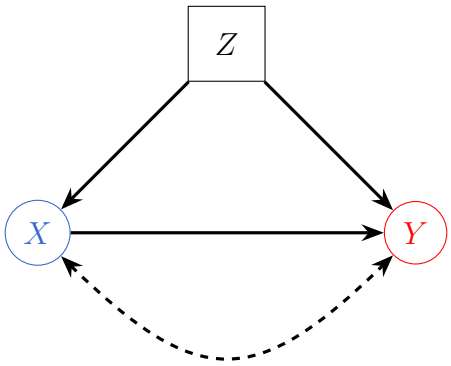


Figure 4.2: Remaining confounding after conditioning on Z.

4.3.3.1 PN Bounds

With data in both observational and experimental settings, conditioning on $\mathbf{Z}$ no longer needs to invoke exogeneity. The bounds of $\text{PN}(\mathbf{z})$ are now:

$$\max\left\{0, \frac{P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z})}{P(x,y|\mathbf{z})}\right\} \leqslant \text{PN}(\mathbf{z}) \leqslant \min\left\{1, \frac{P(y'_{x'}|\mathbf{z}) - P(x',y'|\mathbf{z})}{P(x,y|\mathbf{z})}\right\}. \tag{4.37}$$

Applying the same technique of summing the minimums and maximums of Section 4.3.2, $\text{PN}_{\mathbf{z}}$ is obtained:

$$\text{PN}_{\mathbf{z}} \geqslant \sum_{\mathbf{z}} \max\left\{0, \frac{P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z})}{P(x,y|\mathbf{z})}\right\} \cdot P(z|x,y),$$
$$\text{PN}_{\mathbf{z}} \leqslant \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'_{x'}|\mathbf{z}) - P(x',y'|\mathbf{z})}{P(x,y|\mathbf{z})}\right\} \cdot P(z|x,y).$$

Using lemma 4.3.1, superiority over the Tian-Pearl PN lower bound is *not* obtained when, $\forall i$:

$$0 \leqslant \frac{P(y|\mathbf{z}_i) - P(y_{x'}|\mathbf{z}_i)}{P(x,y|\mathbf{z}_i)},$$
$$P(y_{x'}|\mathbf{z}_i) \leqslant P(y|\mathbf{z}_i),$$

or $\forall i : P(y_{x'}|\mathbf{z}_i) \geqslant P(y|\mathbf{z}_i)$.

Similarly, superiority over the Tian-Pearl PN lower bound is *not* obtained when, $\forall i$:

$$1 \leqslant \frac{P(y'_{x'}|\mathbf{z}) - P(x', y'|\mathbf{z})}{P(x, y|\mathbf{z})},$$

$$P(x, y|\mathbf{z}) \leqslant P(y'_{x'}|\mathbf{z}) - P(x', y'|\mathbf{z}),$$

or $\forall i : P(x, y|\mathbf{z}) \geqslant P(y'_{x'}|\mathbf{z}) - P(x', y'|\mathbf{z})$.

4.3.3.2 P(**benefit**) **Bounds**

The bounds of $P(\text{benefit}|\mathbf{z})$ are:

$$P(\text{benefit}|\mathbf{z}) \geqslant \max \left\{ \begin{array}{l} 0, \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y|\mathbf{z}) \end{array} \right\}, \tag{4.38}$$

$$P(\text{benefit}|\mathbf{z}) \leqslant \min \left\{ \begin{array}{l} P(y_x|\mathbf{z}), \\ P(y'_{x'}|\mathbf{z}), \\ P(x, y|\mathbf{z}) + P(x', y'|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}) + P(x', y|\mathbf{z}) + P(x, y'|\mathbf{z}) \end{array} \right\}. \tag{4.39}$$

Again, the summation of the minimums and maximums technique is applied to bound $P(\text{benefit})_{\mathbf{z}}$:

$$P(\text{benefit})_{\mathbf{z}} \geqslant \sum_{\mathbf{z}} \max \left\{ \begin{array}{l} 0, \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y|\mathbf{z}) \end{array} \right\} \cdot P(\mathbf{z}), \tag{4.40}$$

$$P(\text{benefit})_{\mathbf{z}} \leqslant \sum_{\mathbf{z}} \min \left\{ \begin{array}{l} P(y_x|\mathbf{z}), \\ P(y'_{x'}|\mathbf{z}), \\ P(x, y|\mathbf{z}) + P(x', y'|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}) + P(x', y|\mathbf{z}) + P(x, y'|\mathbf{z}) \end{array} \right\} \cdot P(\mathbf{z}). \tag{4.41}$$

As in the Tian-Pearl $P(\text{benefit})$ bounds of (3.6), the $P(\text{benefit})_{\mathbf{z}}$ bounds have four arguments to the max function and four arguments to the min function. This requires a generalized summation of maximums and minimums lemma:

Lemma 4.3.2. *Given m n-length sequences of values, $\langle\langle x_{1,1}, x_{1,2}, \ldots, x_{1,n}\rangle$, $\langle x_{2,1}, x_{2,2}, \ldots, x_{2,n}\rangle, \ldots, \langle x_{m,1}, x_{m,2}, \ldots, x_{m,n}\rangle\rangle$, the maximum between the summations of each sequence will always be less than or equal to the summation of the maximum between each element of every sequence at the same index:*

$$\max\left\{\sum_i x_{1,i}, \sum_i x_{2,i}, \ldots, \sum_i x_{m,i}\right\} \leqslant \sum_i \max\{x_{1,i}, x_{2,i}, \ldots, x_{m,i}\}. \tag{4.42}$$

This will be equality when $\exists j, \forall k : j \neq k$, each $x_{j,i}$ is greater than or equal to $x_{k,i}$.

Similarly, the minimum between the summations of each sequence will always be greater than or equal to the summation of the minimum between each element of every sequence at the same index:

$$\min\left\{\sum_i x_{1,i}, \sum_i x_{2,i}, \ldots, \sum_i x_{m,i}\right\} \geqslant \sum_i \min\{x_{1,i}, x_{2,i}, \ldots, x_{m,i}\}. \tag{4.43}$$

This will be equality when $\exists j, \forall k : j \neq k$, each $x_{j,i}$ is less than or equal to $x_{k,i}$

The $P(\text{benefit})_{\mathbf{z}}$ lower bound equals the Tian-Pearl $P(\text{benefit})$ lower bound when the same expression in the max function is the maximum for every stratum of $\mathbf{Z}$. Similarly, the $P(\text{benefit})_{\mathbf{z}}$ upper bound equals the Tian-Pearl $P(\text{benefit})$ upper bound when the same expression in the min function is the minimum for every stratum of $\mathbf{Z}$.

4.3.3.3 Graphical Criterion

The graphical criterion for $\mathbf{Z}$ to be advantageous in bounds calculations remains as in Section 4.3.2, with the exception of $\mathbf{Z}$ satisfying the back-door criterion requirement:

- No node in $\mathbf{Z}$ is a descendant of X, unless they are independent of Y_x

4.3.3.4 Pandemic Example

Imagine a terrible pandemic that hits a particular region hard. If unvaccinated, only 37.5% survive. Fortunately, there's a vaccine. While not completely effective, a person has a 75% of survival if vaccinated. Difficult policy decisions need to be made for this vaccine, which is in limited supply. What is the probability of benefiting from this vaccine?

Let X represent vaccination with x being vaccinated and x' being unvaccinated, Y represent survival with y being surviving and y' being succumbing to the pandemic, and Z represent ancestry with z being one ancestral line and z' being the other in this region. The causal graphs in figures 4.1b and 4.1d are examples of this scenario. RCT and observational data reveal $P(z) = P(z') = 0.5$ and the conditional probabilities of Table 4.2.

	Conditioned on z	Conditioned on z'
$P(y_x)$	0.75	0.75
$P(y_{x'})$	0.25	0.5
$P(y\|x)$	0.9	0.6
$P(y\|x')$	0.8	0.5
$P(x)$	0.8	0.25

Table 4.2: Conditional probabilities for pandemic example

The Tian-Pearl bounds are:

$$\max \left\{ \begin{array}{l} 0, \\ 0.75 - 0.375, \\ 0.7025 - 0.375, \\ 0.75 - 0.7025 \end{array} \right\} \leqslant P(\text{benefit}) \leqslant \min \left\{ \begin{array}{l} 0.75, \\ 0.625, \\ 0.435 + 0.2075, \\ 0.75 - 0.375 + 0.2675 + 0.09 \end{array} \right\},$$

$$0.375 \leqslant P(\text{benefit}) \leqslant 0.625.$$

The $P(\text{benefit})_{\mathbf{z}}$ bounds are:

$$P(\text{benefit})_{\mathbf{z}} \geqslant \max\left\{\begin{array}{l} 0, \\ 0.75-0.25, \\ 0.88-0.25, \\ 0.75-0.88 \end{array}\right\} \cdot 0.5 + \max\left\{\begin{array}{l} 0, \\ 0.75-0.5, \\ 0.525-0.5, \\ 0.75-0.525 \end{array}\right\} \cdot 0.5$$

$$= 0.44,$$

$$P(\text{benefit})_{\mathbf{z}} \leqslant \min\left\{\begin{array}{l} 0.75, \\ 0.75, \\ 0.72+0.04, \\ 0.75-0.25+0.16+0.08 \end{array}\right\} \cdot 0.5 + \min\left\{\begin{array}{l} 0.75, \\ 0.5, \\ 0.15+0.375, \\ 0.75-0.5+0.375+0.1 \end{array}\right\} \cdot 0.5$$

$$= 0.62.$$

This example demonstrates Tian-Pearl $P(\text{benefit})$ bounds of $0.375 \leqslant P(\text{benefit}) \leqslant 0.625$ and $P(\text{benefit})_{\mathbf{z}}$ bounds of $0.44 \leqslant P(\text{benefit}) \leqslant 0.62$. This vaccine is more effective than one might have thought looking at the Tian-Pearl bounds. The range decreased as well from 0.25 to 0.18.

4.3.3.5 STAR Real World Example

In Section 3.6, bounds on $P(\text{benefit})$ was estimated, using Tian-Pearl bounds, for the STAR project from Chapter 2. Let us see if we can narrow those bounds further by taking into account the quaternary variable F for first grade math scores (0 to 3 from worst to best). From the DAG in Figure 2.2, it appears that F is an admissible set, satisfying the back-door criterion. However, as noted in Chapter 2, this is a simplified diagram and we do not know if there are other confounders. The lower and upper bounds in Equations (4.40) and (4.41), respectively should be applied to the STAR data laid out in Table 4.3. The notation will use the variable names from the STAR data in Chapter 2, namely outcome variable M (grade 2 math score with m being a high score and m' being a low score), treatment variable S (grade

2 class size with s being a small class size and s' being a regular class size), and confounding variable F.

	$F=0$	$F=1$	$F=2$	$F=3$
$P(m_s\|F)$	0.2743	0.7377	0.9667	1
$P(m_{s'}\|F)$	0.2381	0.6908	0.9822	1
$P(m'_{s'}\|F)$	0.7619	0.3092	0.0178	0
$P(m\|s,F)$	0.3333	0.6863	0.9211	1
$P(m\|s',F)$	0.5909	0.8871	0.9364	0.9444
$P(s\|F)$	0.1154	0.2125	0.1872	0.125
$P(s'\|F)$	0.8846	0.7875	0.8128	0.875
$P(m,s\|F)$	0.04	0.1477	0.1659	0.1429
$P(m,s'\|F)$	0.52	0.6962	0.7678	0.8095
$P(m',s\|F)$	0.08	0.0675	0.0142	0
$P(m',s'\|F)$	0.36	0.0886	0.0521	0.0476
$P(m\|F)$	0.56	0.8439	0.9336	0.9524
$P(F)$	0.1011	0.5303	0.3375	0.0311

Table 4.3: Conditional probabilities for STAR real world example

$$
\begin{aligned}
P(\text{benefit})_{\mathbf{z}} &\geqslant \max\{0, 0.0362, 0.3219, -0.2857\} \cdot 0.1011 \\
&\quad + \max\{0, 0.0469, 0.1531, -0.1062\} \cdot 0.5303 \\
&\quad + \max\{0, -0.0155, -0.0486, 0.0331\} \cdot 0.3375 \\
&\quad + \max\{0, 0, -0.0476, 0.0476\} \cdot 0.0311 \\
&= 0.3219 \cdot 0.1011 + 0.1531 \cdot 0.5303 + 0.0331 \cdot 0.3375 + 0.0476 \cdot 0.0311 \\
&\approx 0.1264, \\
P(\text{benefit})_{\mathbf{z}} &\leqslant \min\{0.2743, 0.7619, 0.4, 0.6362\} \cdot 0.1011 \\
&\quad + \min\{0.7377, 0.3092, 0.2363, 0.8106\} \cdot 0.5303 \\
&\quad + \min\{0.9667, 0.0178, 0.218, 0.7665\} \cdot 0.3375 \\
&\quad + \min\{1, 0, 0.1905, 0.8095\} \cdot 0.0311 \\
&= .2743 \cdot 0.1011 + 0.2363 \cdot 0.5303 + 0.0178 \cdot 0.3375 + 0 \cdot 0.0311 \\
&\approx 0.159.
\end{aligned}
$$

Let us compare these results with the Tian-Pearl bound results in Equation (3.16):

$$
\begin{aligned}
0.1264 \leqslant P(\text{benefit}) \leqslant 0.235, \\
0.1264 \leqslant P(\text{benefit})_{\mathbf{z}} \leqslant 0.159.
\end{aligned}
$$

We see that the lower bound was unchanged. However, the upper bound has come down considerably. We are now significantly more precise in our estimate of the proportion of students that benefit from smaller class sizes.

4.3.4 Additional Information Paradox

Let us revisit the pandemic example of Section 4.3.3.4 to mirror the discussion in [MLP22]. This time there is only RCT data and the population is evenly split by ancestry, referenced in conditional probabilities Table 4.4.

	Conditioned on z	Conditioned on z'
$P(y_x)$	0.75	0.25
$P(y_{x'})$	0.2	0.6

Table 4.4: Conditional probabilities for pandemic example RCT

Four different bounds can be calculated for $P(\text{benefit})$:

$$\begin{aligned}
\text{Tian-Pearl} &\implies 0.1 \leqslant P(\text{benefit}) \leqslant 0.5 \\
\text{Covariate-improved} &\implies 0.275 \leqslant P(\text{benefit})_z \leqslant 0.5 \\
\text{Person has ancestry } z &\implies 0.55 \leqslant P(\text{benefit}|z) \leqslant 0.75 \\
\text{Person has ancestry } z' &\implies 0 \leqslant P(\text{benefit}|z') \leqslant 0.25
\end{aligned}$$

As expected, bounds on $P(\text{benefit})_z$ are narrower than the Tian-Pearl $P(\text{benefit})$ bounds. Surprisingly, if a person knows their ancestry, then their $P(\text{benefit})$ bounds are completely outside the $P(\text{benefit})_z$ bounds. Basically, knowing your ancestry gives you very different, not necessarily narrower, $P(\text{benefit})$ bounds than not knowing your ancestry.

The paradox is that additional information of ancestral knowledge violates the heuristic that *additional information* should narrow the bounds or keep them the same. If someone's ancestry is unknown, the probability they benefit from this vaccine is between 0.275 and 0.5. Once the additional information is acquired that the person is of ancestry z, the probability they benefit from this treatment becomes between 0.55 and 0.75. It seems their probability of benefiting never really was between 0.275 and 0.5.

The reason for this seeming inconsistency is that there are two different questions. Without knowing the ancestry, the question is, "what is the probability of benefiting for a person regardless of ancestry?" With knowing the ancestry, the question becomes, "what is the probability of benefiting for a person of ancestry Z?" The additional information of the person's ancestry didn't help the first question and the second question is not answerable

without the additional information.

The following example will illuminate the reasons for this phenomenon [Pea09, page 296]. Let the covariate Z stand for the outcome of a fair coin toss, so $P(Z = \text{heads}) = 0.5$. Without knowing what treatment X and success Y represent, let $P(y_x) = P(y_{x'}) = 0.5$. The remaining probabilities are in Table 4.5.

	Conditioned on heads	Conditioned on tails
$P(y_x)$	1	0
$P(y_{x'})$	0	1

Table 4.5: Conditional probabilities for coin toss example

Tian-Pearl $P(\text{benefit})$ bounds are $0 \leqslant P(\text{benefit}) \leqslant 0.5$ and $P(\text{benefit})_z$ bounds are $0.5 \leqslant P(\text{benefit}) \leqslant 0.5$ or $P(\text{benefit})_z = 0.5$.

Now, let us uncover the functional mechanism, x represents betting \$1 on heads, x' represents betting \$1 on tails, y represents winning \$1, and y' represents losing \$1. It should now be clear why $P(y_x) = P(y_{x'}) = 0.5$. Without knowing the coin toss result, the odds of winning \$1 are 50/50 whether you bet on heads or tails. The $P(\text{benefit})$ is also 0.5 because benefiting from betting on heads is true only when the coin toss was heads and the coin toss is heads 50% of the time.

This brings us back to the $P(\text{benefit})$ bounds when we have the additional information of what the coin toss result was. If we know the coin toss resulted in heads, then the probability of benefiting from betting on heads is 100%. Similarly, if we know the coin toss resulted in tails, then the probability of benefiting from betting on heads is 0%. In other words, $P(\text{benefit}|\text{heads}) = 1$ and $P(\text{benefit}|\text{tails}) = 0$. If the coin toss is heads, winning only happens when betting on heads. Even though the bounds are completely different when we provided with the very useful additional information of the coin toss, there is clearly no contradiction here. There was a 50% probability of benefiting from betting on heads when

we didn't know the coin toss result and a 100% probability of benefiting from betting on heads when we knew the coin toss resulted in heads. We were asking two separate questions. The first question was, "what is the probability of benefiting without knowing the coin toss result?" The second question was, "what is the probability of benefiting for a coin toss of heads?"

4.3.5 Practical Usage

Knowledge of a causal structure enables narrower PN, PS, and P(benefit) bounds to be estimated compared with the tight bounds of Tian and Pearl which were derived without such knowledge. This mechanism can be used whenever the graphical criterion of sections 4.3.2.3 and 4.3.3.3 are satisfied. These are weighted averages of the $\mathbf{Z}$-specific probabilities of causation. If an individual's $\mathbf{Z}$ values are known, the bounds of PN($\mathbf{z}$) and P(benefit|$\mathbf{z}$) in equations 4.23, 4.32, 4.37, 4.38, and 4.39 should be consulted.

Examples in sections 4.3.2.4 and 4.3.3.4 showcase the situation where data for a covariate is available on the population, but not on the individual we are trying to answer the query for. This is important to note. If an individual knows their covariate set values, then the data should be conditioned on and Tian-Pearl bounds should be consulted. Personal decision making only benefits from the techniques in this chapter when population data is known, but individual covariate data is unknown.

Another scenario where covariates can improve bounds using the techniques presented in this chapter is when covariate data for an individual is known, but the sample size in that stratum is too small. For example, natural hair color affects effectiveness of some medication. The person of interest has red hair. There's only one other person that has taken the medication and had red hair. It is not possible to get an accurate P(benefit) estimate, so a weighted average of P(benefit|$\mathbf{z}$) will be most accurate and informative. This is analogous to using ATE instead of CATE because of too little data in the conditioning set.

4.4 Leveraging Mediation Data

Section 4.3 discussed evaluation of PN, PS, and P(benefit) using a set of covariates, as long as that set did not include any descendants of the treatment. Descendants were allowed, though not helpful, if they were independent of the potential outcome given a particular treatment. These assumptions exclude mediators, variables within a causal path from treatment to outcome. However, mediators lend themselves well to practical usage of narrowing bounds on probabilities of causation.

Section 4.3.5 considered situations which work well for incorporating covariates to narrow bounds. The scenario most conducive is when population data is available and individual covariate data is unavailable. Unfortunately, this is not always the situation confronting us. We typically either have covariate data on both the population and the individual under consideration or we lack covariate data for both the population and the individual.

However, we frequently need to know the probability of a cause for an individual where mediator data on the population is available, but mediator data on that individual is not. This is because a mediator is a descendant of the treatment, which makes it necessarily post-treatment. In the case of personal decision making, it is critical to know the benefit of treatment, whether the treatment will provide a favorable outcome and no treatment will yield an unfavorable outcome. This P(benefit) query is posed before treatment is taken, when post-treatment data availability would be rare or impossible.

4.4.1 Pure Mediator

This section will examine mediator sets $\mathbf{M}$ with the simplifying assumption that binary treatment X affects binary outcome Y only through $\mathbf{M}$. In the context of IVs, this is known as the exclusion restriction [LS18], where X plays the role of an instrument. Let us repurpose exclusion restriction here. Let M be referred to as a pure mediator when X affects Y only through M. In Section 4.4.2, this assumption will be relaxed. Figure 4.3 depicts example

causal graphs of this scenario.

(a) Simple pure mediator M (b) X, Y, and mediator M pairwise confounded

Figure 4.3: Mediators where X affects Y only through M.

4.4.1.1 P(benefit) Bounds

The following analysis will start with the P(benefit) instead of starting with the PN as in previous sections. The reason is that the PN can be easily derived from the P(benefit) with an additional assumption.

For simplicity, M is a single binary mediator with values m and m'. Section 4.4.1.4 will generalize the following bounds with non-binary mediator sets. Intuition around using mediators to narrow bounds will be introduced, followed by formulas and graphical criteria.

The probability of benefiting, $P(\text{benefit}) = P(y_x, y'_{x'})$, is the probability of recovery had the individual been treated and non-recovery had the individual not been treated. This can happen through a binary pure mediator in two ways. The first is that X benefits M and M benefits Y. In other words, $M = m$ upon treatment and $M = m'$ upon no treatment. And $Y = y$ upon m and $Y = y'$ upon m'. The probability of X benefiting M is simply the P(benefit) for X and M, $P(m_x, m'_{x'})$. Similarly, the probability of M benefiting Y is the P(benefit) for M and Y, $P(y_m, y'_{m'})$. The quantity of interest, P(benefit), is the probability of a unit or individual being a benefiter from X to Y. This happens when units or individuals who are benefiters from X to M and also benefiters from M to Y. Let us call these individuals double-benefiters.

The second way X can benefit Y is when X *harms* M and M harms Y. In other words,

a unit or individual is harmed from X to M and also harmed from M to Y. Let us call these individuals double-harmed.

Let PN_m and $P(\text{benefit})_m$ be the PN and $P(\text{benefit})$, respectively, when evaluating with the mediator M. The probability of an individual being a double-benefiter or a double-harmed is the $P(\text{benefit})$:

$$\begin{aligned} P(\text{benefit})_m &= P(y_x, y'_{x'}) \\ &= P(y_m, y'_{m'}, m_x, m'_{x'}) + P(y'_m, y_{m'}, m'_x, m_{x'}). \end{aligned} \tag{4.44}$$

In addition to the assumption of X affecting Y only through M, the second assumption to be made is $(Y_m, Y_{m'}) \perp\!\!\!\perp (M_x, M_{x'})$. This allows splitting (4.44):

$$P(\text{benefit})_m = P(y_m, y'_{m'}) \cdot P(m_x, m'_{x'}) + P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}). \tag{4.45}$$

The Fréchet inequalities for two events in Equation (3.17) will be used in the bound-narrowing techniques below.

Using the following technique, the $P(\text{benefit})_m$ upper bound can sometimes be smaller than the Tian-Pearl upper bound. However, the $P(\text{benefit})_m$ lower bound receives no such advantage. To see why, let us apply the left side of (3.17) to the probability of being a double-benefiter:

$$\begin{aligned} P(\text{double-benefiter}) &= P(y_m, y'_{m'}, m_x, m'_{x'}) \\ &= P(y_m, y'_{m'}) \cdot P(m_x, m'_{x'}) \\ &\geqslant \max\{0, P(y_m) - P(y_{m'})\} \cdot \max\{0, P(m_x) - P(m_{x'})\} \\ &= \max\{0, [P(y_m) - P(y_{m'})] \cdot [P(m_x) - P(m_{x'})]\}. \end{aligned} \tag{4.46}$$

Next, let us apply the left side of (3.17) to the probability of being a double-harmed:

$$\begin{aligned} P(\text{double-harmed}) &= P(y'_m, y_{m'}, m'_x, m_{x'}) \\ &= P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}) \\ &\geqslant \max\{0, P(y'_m) - P(y'_{m'})\} \cdot \max\{0, P(m'_x) - P(m'_{x'})\} \\ &= \max\{0, [P(y_m) - P(y_{m'})] \cdot [P(m_x) - P(m_{x'})]\}. \end{aligned} \tag{4.47}$$

Equations (4.46) and (4.47) are the same. Lemma 4.3.1 tells us that we can expect no advantage for $P(\text{benefit})_m$ over the Tian-Pearl $P(\text{benefit})$ in this case.

The upper bound, on the other hand, can be lowered. As before, let us start with the right side of (3.17):

$$\begin{aligned} P(\text{double-benefiter}) &= P(y_m, y'_{m'}, m_x, m'_{x'}) \\ &= P(y_m, y'_{m'}) \cdot P(m_x, m'_{x'}) \\ &\leqslant \min\{P(y_m), P(y'_{m'})\} \cdot \min\{P(m_x), P(m'_{x'})\} \\ &= \begin{cases} P(y_m) \cdot P(m_x), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\ P(y_m) \cdot P(m'_{x'}), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}), \\ P(y'_{m'}) \cdot P(m_x), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\ P(y'_{m'}) \cdot P(m'_{x'}), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}), \end{cases} \\ P(\text{double-harmed}) &= P(y'_m, y_{m'}, m'_x, m_{x'}) \\ &= P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}) \\ &\leqslant \min\{P(y_{m'}), P(y'_m)\} \cdot \min\{P(m_{x'}), P(m'_x)\} \\ &= \begin{cases} P(y_{m'}) \cdot P(m_{x'}), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\ P(y_{m'}) \cdot P(m'_x), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}), \\ P(y'_m) \cdot P(m_{x'}), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\ P(y'_m) \cdot P(m'_x), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}). \end{cases} \end{aligned} \tag{4.48}$$

Combining $P(\text{double-benefiter})+P(\text{double-harmed})$ yields the upper bound of $P(\text{benefit})_m$:

$$P(\text{benefit})_m \leqslant \min \left\{ \begin{array}{l} P(y_m) \cdot P(m_x) + P(y_{m'}) \cdot P(m_{x'}), \\ P(y_m) \cdot P(m'_{x'}) + P(y_{m'}) \cdot P(m'_x), \\ P(y'_{m'}) \cdot P(m_x) + P(y'_m) \cdot P(m_{x'}), \\ P(y'_{m'}) \cdot P(m'_{x'}) + P(y'_m) \cdot P(m'_x) \end{array} \right\}. \tag{4.49}$$

If there is no pairwise confounding between X, M, and Y, as in Figure 4.3a, then this simplifies to observational probabilities:

$$P(\text{benefit})_m \leqslant \min \left\{ \begin{array}{l} P(y|m) \cdot P(m|x) + P(y|m') \cdot P(m|x'), \\ P(y|m) \cdot P(m'|x') + P(y|m') \cdot P(m'|x), \\ P(y'|m') \cdot P(m|x) + P(y'|m) \cdot P(m|x'), \\ P(y'|m') \cdot P(m'|x') + P(y'|m) \cdot P(m'|x) \end{array} \right\}. \tag{4.50}$$

This upper bound for $P(\text{benefit})_m$ is sometimes worse than Tian-Pearl's $P(\text{benefit})$ upper bound. So, the overall upper bound is:

$$P(\text{benefit})_m \leqslant \min \left\{ \begin{array}{l} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + P(x',y) \ + P(x,y'), \\ P(y_m) \cdot P(m_x) + P(y_{m'}) \cdot P(m_{x'}), \\ P(y_m) \cdot P(m'_{x'}) + P(y_{m'}) \cdot P(m'_x), \\ P(y'_{m'}) \cdot P(m_x) + P(y'_m) \cdot P(m_{x'}), \\ P(y'_{m'}) \cdot P(m'_{x'}) + P(y'_m) \cdot P(m'_x) \end{array} \right\}. \tag{4.51}$$

The third and fourth arguments are eliminated if observational data is unavailable.

4.4.1.2 PN Bounds

Under strong exogeneity [TP00], $P(\text{benefit})$ and PN are related with:

$$\text{PN} = \frac{P(\text{benefit})}{P(y|x)}.$$

This means the bounds for PN_m with the models depicted in figures 4.3a and 4.4 are simply the bounds for $P(\text{benefit})_m$ divided by $P(y|x)$.

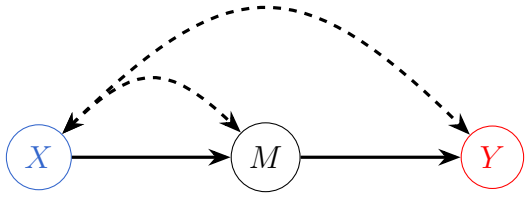


Figure 4.4: Pure mediator with $X \to M$ and $X \to Y$ confounding

However, strong exogeneity does not hold in the graphs of figures 4.3b and 4.4. The reasons will be seen in Section 4.4.1.3. Strong exogeneity does hold in the graph of Figure 4.5.

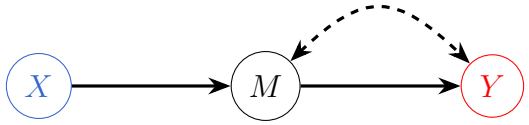


Figure 4.5: Pure mediator with $M \to Y$ confounding

The same bounds were obtained in [DMM17] for PN_m on the simple pure mediator of Figure 4.3a.

4.4.1.3 Graphical Criterion

Two constraints were declared in the derivation to obtain potentially smaller upper bounds on $P(\text{benefit})_m$:

- Binary treatment X affects binary outcome Y only through $\mathbf{M}$ (exclusion restriction)
- $(Y_m, Y_{m'}) \perp\!\!\!\perp (M_x, M_{x'})$

The first constraint is easy to visualize with a conventional DAG. Simply ensure all directed unblocked paths from X to Y contain $\mathbf{M}$.

The second constraint is impossible to visualize and even difficult to intuit with conventional DAGs due to its counterfactual terms. Alternative graphical methods have been devised to process and visualize counterfactual criterion like this, such as Single-World Intervention Graphs (SWIG) [RR13], Twin Networks [BP94], and the Parallel Worlds graph [SP12, SP08]. A SWIG is appropriate for this scenario as it requires a single hypothetical world, one in which X is either x or x' and M is either m or m'. Partial mediators of Section 4.4.2 will require multiple hypothetical worlds.

This SWIG is drawn in Figure 4.6 for Figure 4.3b. The X node is split into its random component, X, and its fixed component, x. Random component parts inherit incoming edges, while the fixed component parts inherit outgoing edges. Because x then points to M, the random component of M becomes M_x.

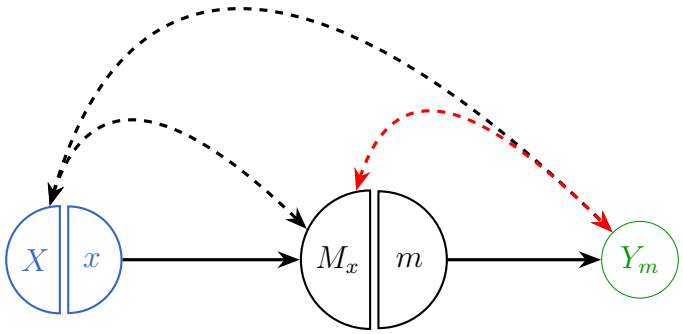


Figure 4.6: Pure mediator SWIG with pairwise confounding

Notice that Y_m and M_x are d-separated when the red bidirectional dashed arrow between them is removed.

The graphical criterion for PN_m has an additional constraint:

- $Y_x \perp\!\!\!\perp X$

The SWIG in Figure 4.7 visualizes this counterfactual independency.

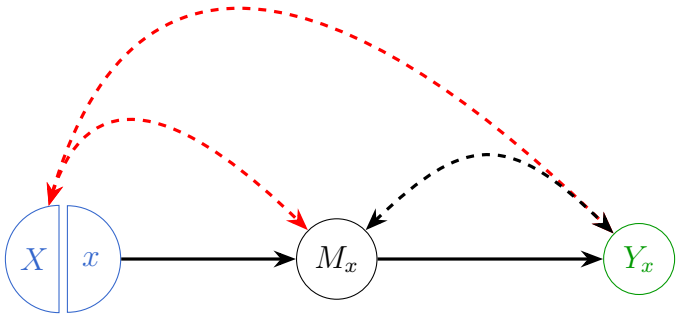


Figure 4.7: Pure mediator SWIG with $Y_x \perp\!\!\!\perp X$ violations in red

It is clear from this SWIG that the X and Y cannot have any shared ancestors [GP07] (represented as the red bidirectional dashed arrows) and neither can X and M have any shared ancestors in order for $Y_x \perp\!\!\!\perp X$. To use the techniques in this section to narrow bounds on PN, there also cannot be any shared ancestors between M and Y.

4.4.1.4 Non-binary

The $P(\text{benefit})_m$ upper bound of Section 4.4.1.1 can be generalized to non-binary mediator sets:

$$\begin{aligned} P(\text{benefit})_{\mathbf{m}} &= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_{\mathbf{m}_i}, y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \\ &\leqslant \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_{\mathbf{m}_i}), P(y'_{\mathbf{m}_j})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} \qquad (4.52) \\ &= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}: i<j} \min \left\{ \begin{array}{l} P(y_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{i_x}) + P(y_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_{x'}}), \\ P(y_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{j_{x'}}) + P(y_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{j_x}), \\ P(y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_x}) + P(y'_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{i_{x'}}), \\ P(y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{j_{x'}}) + P(y'_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{j_x}) \end{array} \right\}. \qquad (4.53) \end{aligned}$$

Each term in the summation of (4.53) comprises two terms in the summation of (4.52). This is because Equation (4.49) works for each pair of values in $\mathbf{M}$, where $m = \mathbf{m}_i \in \mathbf{M}$,

$m' = \mathbf{m}_j \in \mathbf{M}$, and $i \neq j$. The constraint of $i < j$ in the summation ensures these terms aren't added twice.

4.4.1.5 Example

Imagine a vaccine that protects from a disease purely by producing antibodies. Let x and x' represent getting and not getting the vaccine, respectively, m and m' represent high antibody count and low antibody count, respectively, and y and y' represent avoiding and acquiring the disease, respectively. Researchers have consensus that high antibody count is only possible through the vaccine or, surprisingly, completely at random. And acquiring the disease depends completely on antibody count and randomness. This implies there is no pairwise confounding between X, Y, and M. The casual graph is depicted in Figure 4.3a. The following data are collected:

$$P(y|m) = 0.5,$$
$$P(y|m') = 0.5,$$
$$P(m|x) = 0.1,$$
$$P(m|x') = 0.1.$$

Comparing Tian-Pearl's $P(\text{benefit})$ with $P(\text{benefit})_m$ is straightforward:

$$\begin{aligned}
P(\text{benefit}) &\leqslant \min \left\{ \begin{array}{l} P(y|m) \cdot P(m|x) + P(y|m') \cdot P(m'|x), \\ P(y'|m) \cdot P(m|x') + P(y'|m') \cdot P(m'|x') \end{array} \right\} \\
&= \min \left\{ \begin{array}{l} 0.5 \cdot 0.1 + 0.5 \cdot 0.9, \\ 0.5 \cdot 0.1 + 0.5 \cdot 0.9 \end{array} \right\} \\
&= 0.5, \\
P(\text{benefit})_m &\leqslant \min \left\{ \begin{array}{l} P(y|m) \cdot P(m|x) + P(y|m') \cdot P(m|x'), \\ P(y|m) \cdot P(m'|x') + P(y|m') \cdot P(m'|x), \\ P(y'|m') \cdot P(m|x) + P(y'|m) \cdot P(m|x'), \\ P(y'|m') \cdot P(m'|x') + P(y'|m) \cdot P(m'|x) \end{array} \right\} \\
&= \min \left\{ \begin{array}{l} 0.5 \cdot 0.1 + 0.5 \cdot 0.1, \\ 0.5 \cdot 0.9 + 0.5 \cdot 0.9, \\ 0.5 \cdot 0.1 + 0.5 \cdot 0.1, \\ 0.5 \cdot 0.9 + 0.5 \cdot 0.9 \end{array} \right\} \\
&= 0.1.
\end{aligned}$$

The $P(\text{benefit})_m$ upper bound is significantly smaller than what the Tian-Pearl upper bound provides, 0.1 versus 0.5. This means the benefit to taking the vaccine is at best 10%.

Since Figure 4.3a satisfies strong exogeneity:

$$\begin{aligned}
P(y|x) &= P(y|m) \cdot P(m|x) + P(y|m') \cdot P(m'|x) \\
&= 0.5.
\end{aligned}$$

Therefore, the Tian-Pearl upper bound is $\text{PN} \leqslant \frac{0.5}{0.5} = 1$ and $\text{PN}_m \leqslant \frac{0.1}{0.5} = 0.2$. The Tian-Pearl bounds offered no information on the probability that the vaccine was necessary to avoid acquiring the disease for a person who took the vaccine and avoided the disease. While PN_m was very informative in that there was a maximum of 20% chance the vaccine was necessary.

4.4.2 Partial Mediator

With partial mediation, the requirement that the treatment X affects the outcome Y only through a mediator M is relaxed. An example partial mediator is shown in Figure 4.8.

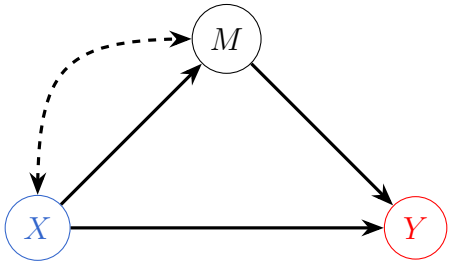


Figure 4.8: Partial mediator M with $X \to M$ confounding

4.4.2.1 P(benefit) Bounds

The following derivation for $P(\text{benefit})_m$ uses three assumptions to obtain measurable probabilities:

- $Y_x \perp\!\!\!\perp M_{x'} \mid M_x$
- $Y_{x'} \perp\!\!\!\perp M_x \mid M_{x'}$
- $Y_x \perp\!\!\!\perp X \mid M_x$

This derivation is equivalent to the proof of theorem 6 in [MLP22]:

$$
\begin{aligned}
P(\text{benefit})_{\mathbf{m}} &= P(y_x, y'_{x'}) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_x, y'_{x'}, \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) && (4.54) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_x, y'_{x'} | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \cdot P(\mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \\
&\leqslant \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}), P(y'_{x'} | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} \\
& && (4.55) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}), P(y'_{x'} | \mathbf{m}_{j_{x'}})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} && (4.56) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}, x), P(y'_{x'} | \mathbf{m}_{j_{x'}}, x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} && (4.57) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y | \mathbf{m}_i, x), P(y' | \mathbf{m}_j, x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\}.
\end{aligned}
$$

Equation (4.54) is a simple application of total probability, Equation (4.55) splits the conditional $P(\text{benefit})$ using the Fréchet upper bound, Equation (4.56) relies on the assumptions declared above, $Y_x \perp\!\!\!\perp M_{x'} \mid M_x$ and $Y_{x'} \perp\!\!\!\perp M_x \mid M_{x'}$, and Equation (4.57) relies on the above assumption, $Y_x \perp\!\!\!\perp X \mid M_x$.

Note that if there is no confounding between X and M, then only observational data is used in this $P(\text{benefit})_m$ upper bound.

Just like the pure mediator case of Section 4.4.1, sometimes Tian-Pearl upper bounds are

smaller than this $P(\text{benefit})_{\mathbf{m}}$ upper bound. So, the overall upper bound is:

$$P(\text{benefit})_m \leqslant \min \left\{ \begin{array}{l} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + P(x',y) \; + P(x,y'), \\ \sum\limits_{\mathbf{m}_i,\mathbf{m}_j \in \mathbf{M}} \min\{P(y|\mathbf{m}_i,x), P(y'|\mathbf{m}_j,x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} \end{array} \right\}, \tag{4.58}$$

with the third and fourth arguments eliminated if observational data is unavailable.

4.4.2.2 PN Bounds

As in Section 4.4.1.2, strong exogeneity allows easily computing PN_m from $P(\text{benefit})_m$ by dividing by $P(y|x)$. The $X \rightarrow M$ confounding of Figure 4.8 does not satisfy strong exogeneity, while the graph in Figure 4.9 does.

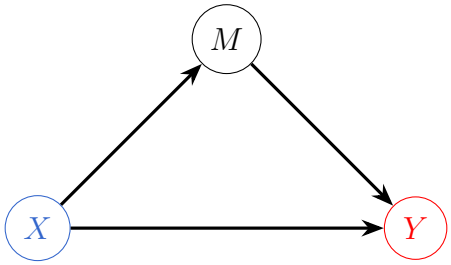


Figure 4.9: Partial mediator M with no confounding among any variable pair

4.4.2.3 Graphical Criterion

The following criterion was stated for the derivation of $P(\text{benefit})_m$ in this section:

- $Y_x \perp\!\!\!\perp M_{x'} \mid M_x$
- $Y_{x'} \perp\!\!\!\perp M_x \mid M_{x'}$

- $Y_x \perp\!\!\!\perp X \mid M_x$

The SWIGs in the figures of Section 4.4.1.3 are no longer sufficient to visualize and verify these assumptions. The counterfactual terms of the first two assumptions involve multiple hypothetical worlds. Figure 4.8 is drawn as a Parallel Worlds graph in Figure 4.10.

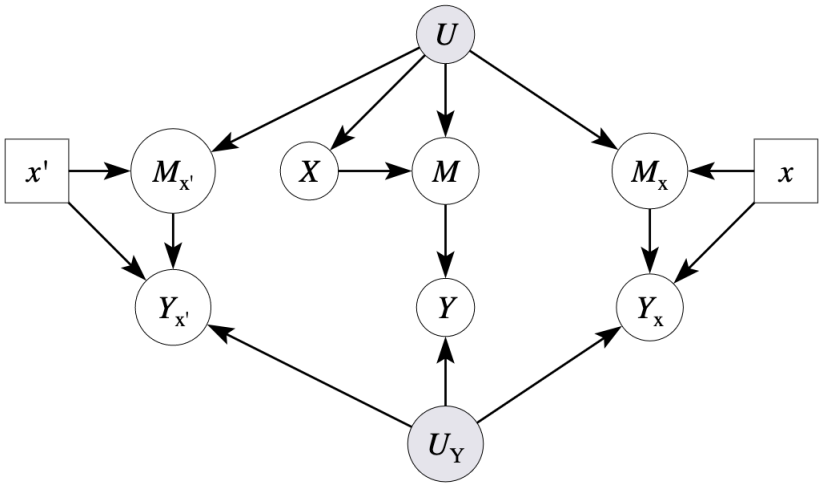


Figure 4.10: Partial mediator Parallel Worlds graph

Confounding between two variables must now be drawn as confounding between those two variables in all worlds. For example, confounding between X and M becomes confounding between X, M, M_x, and $M_{x'}$. Square boxes enclosing x and x' indicate they are held fixed. It can then be seen that Y_x is d-separated from $M_{x'}$ given M_x, symmetrically $Y_{x'}$ is d-separated from M_x given $M_{x'}$, and Y_x is d-separated from X given M_x.

4.5 Leveraging Combinations of Covariates

Sections 4.3 and 4.4 described how to apply sets of covariates and how to apply sets of mediators, respectively, under different criteria, to narrow bounds on PN, PS, and P(benefit).

This section will briefly analyze how to overcome criterion violations with mediators and combining covariates and mediators.

4.5.1 Mediator with Confounding

Figure 4.11 shows a causal graph where the criterion for bounds on $P(\text{benefit})_{\mathbf{m}}$ using mediator M is not satisfied. In particular, $(Y_m, Y_{m'}) \perp\!\!\!\perp (M_x, M_{x'})$ is violated by the confounder Z.

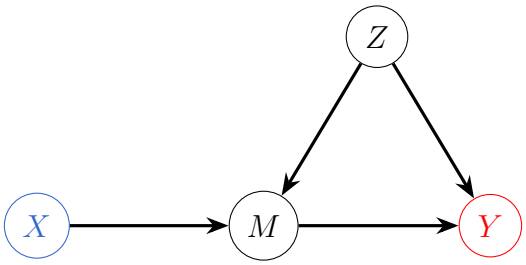


Figure 4.11: Pure mediator with $M \to Y$ confounded by Z

4.5.1.1 Pure Mediator

Since blocking on Z allows $(Y_m, Y_{m'}) \perp\!\!\!\perp (M_x, M_{x'})$, the upper bound on $P(\text{benefit})_{\mathbf{m}}$ can be computed for each stratum of Z. The final result is a weighted average on the upper bounds by $P(z)$:

$$P(\text{benefit})_{\mathbf{m}} \leqslant \mathbb{E}_Z \left[\sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}: i<j} \min \left\{ \begin{array}{l} P(y_{\mathbf{m}_i}|z) \cdot P(\mathbf{m}_{i_x}|z) + P(y_{\mathbf{m}_j}|z) \cdot P(\mathbf{m}_{i_{x'}}|z), \\ P(y_{\mathbf{m}_i}|z) \cdot P(\mathbf{m}_{j_{x'}}|z) + P(y_{\mathbf{m}_j}|z) \cdot P(\mathbf{m}_{j_x}|z), \\ P(y'_{\mathbf{m}_j}|z) \cdot P(\mathbf{m}_{i_x}|z) + P(y'_{\mathbf{m}_i}|z) \cdot P(\mathbf{m}_{i_{x'}}|z), \\ P(y'_{\mathbf{m}_j}|z) \cdot P(\mathbf{m}_{j_{x'}}|z) + P(y'_{\mathbf{m}_i}|z) \cdot P(\mathbf{m}_{j_x}|z) \end{array} \right\} \right].$$

This can be compared with the Tian-Pearl upper bound to find the smallest upper bound.

4.5.1.2 Partial Mediator

Similarly, Figure 4.12 shows a partial mediator where the $P(\text{benefit})_{\mathbf{m}}$ upper bound cannot easily be computed due to the violations of $Y_x \perp\!\!\!\perp M_{x'} \mid M_x$ and $Y_{x'} \perp\!\!\!\perp M_x \mid M_{x'}$.

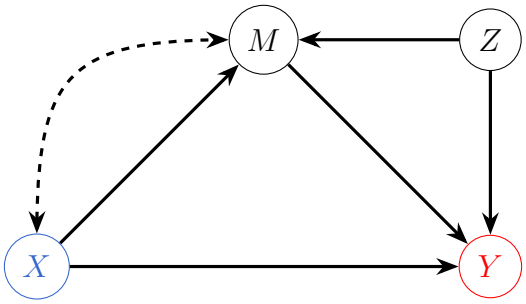


Figure 4.12: Pure mediator with $M \to Y$ confounded by Z

Taking the weighted average of the upper bound on $P(\text{benefit})_{\mathbf{m}}$ for each stratum of Z yields:

$$P(\text{benefit})_{\mathbf{m}} \leqslant \mathbb{E}_Z \left[\sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y|\mathbf{m}_i, x, z), P(y'|\mathbf{m}_j, x', z)\} \cdot \min\{P(\mathbf{m}_{i_x}|z), P(\mathbf{m}_{j_{x'}}|z)\} \right].$$

This can be compared with the Tian-Pearl upper bound to find the smallest upper bound.

4.5.2 Covariates and Mediators

Figure 4.13 presents a causal graph with the possibility of deriving bounds on $P(\text{benefit})_{\mathbf{m}}$ using the mediator M or the covariate Z.

Whenever multiple possibilities exist for computing bounds, simply use the largest lower bound and smallest upper bound.

4.6 Summary

This chapter analyzed and presented methods of computing narrower bounds on probabilities of causation, PN, PS, and $P(\text{benefit})$. It demonstrates how significant narrowing of bounds

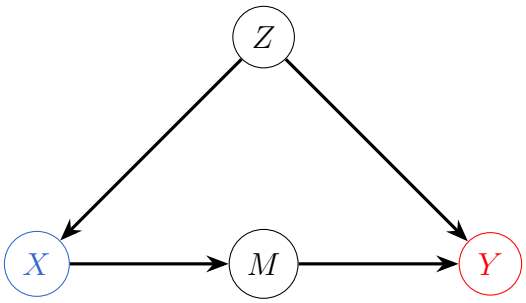


Figure 4.13: Pure mediator M with $X \to Y$ confounded by Z

on probabilities of causation can be attained and how this impacts decision making at every level and in almost every discipline.

Given the fertile ground for improvement and their significant impact, the question arises, why hasn't there been more formal research in this area? One possibility is that researchers often care about effects instead of causes. Hypotheses, core to the scientific method, are typically in the form of EoC. This can possibly account for the reason that more effort has gone into the development of curricula, pedagogy, tools, and software around EoC. This in turn reinforced the focus on EoC – that is what they were taught.

Another reason for the lopsided emphasis on EoC, alluded to in Section 4.1, is the difficulty or rarity of obtaining point estimates or sufficiently narrow bounds on CoE. Section 4.2.5 reviewed how PN, PS, and P(benefit) point estimates can be obtained if the strong assumption of monotonicity holds. However, even when monotonicity holds, it can be a challenge to be convinced of it and monotonicity is generally untestable. Point estimates, or any counterfactual term, can be computed if the SCM is known. Unfortunately, this knowledge is rare. Tian and Pearl [TP00] derived bounds on PN, PS, and P(benefit) and proved their tightness. The inability of improving on these bounds to result in sufficiently narrow bounds may have contributed to the lack of research interest in CoE. We should be reminded, "We learned from Simpson's Paradox that certain decisions cannot be made on the basis of data alone, but instead depend on the story behind the data" [PGJ16, page 24]. Population data alone can never improve bounds on CoE, however, the story behind the data can.

Valuable CoE work, like Li and Pearl's *Unit Selection based on Counterfactual Logic* [LP19], that depend on P(benefit), P(harm), P(immunity), and P(doom), is significantly enhanced with better accuracy by narrowing these bounds.

CHAPTER 5

Leveraging Concurrent-Controlled RCT Data

5.1 Background

This chapter introduces an important restriction on data obtained from RCTs. It is often assumed that RCTs provide estimates of two causal effects, $P(y_x)$ and $P(y_{x'})$, standing for the probability of the outcome Y under treatment and control, respectively. In medical practices, however, these two quantities are rarely reported separately; only their difference ATE = $P(y_x) - P(y_{x'})$ is measured, estimated, and reported. The reason is that the individual effects, $P(y_x)$ and $P(y_{x'})$, are suspect of contamination by selection bias and placebo effects. These two imperfections are presumed to cancel out by a method called "Concurrent Control" [Sen10] in which subjects in both treatment and control arms are measured simultaneously and only the average difference, ATE, is counted.

This chapter establishes bounds on P(benefit), P(harm), P(immunity), and P(doom) under the restriction that RCTs provide only an assessment of ATE, not of the individual causal effects $P(y_x)$ and $P(y_{x'})$. It will be shown that the new restriction, though leading to less narrow bounds than obtained with both causal effects, still permits the extraction of meaningful information on individual benefit, harm, immunity, and doom and, when combined with observational data, can be extremely valuable in personalized decision making.

5.2 P(benefit) Bounds

With just the ATE, P(benefit) is bounded as:

$$\max\{0, \text{ATE}\} \leqslant P(\text{benefit}) \leqslant \min\{1, \text{ATE} + 1\}. \tag{5.1}$$

The lower bound is above zero when ATE is positive, and the upper bound is lower than 1 when ATE is negative.

When we combine ATE with observational data, the lower bound remains the same, but the upper bound changes to yield:

$$\max\{0, \text{ATE}\} \leqslant P(\text{benefit}) \leqslant \min \left\{ \begin{array}{r} P(x,y) + P(x',y'), \\ \text{ATE} + P(x,y') + P(x',y) \end{array} \right\}. \tag{5.2}$$

The upper bound in (5.2), is always lower than (or equal to) the one in (5.1), because both $P(x,y) + P(x',y') \leqslant 1$ and $P(x,y') + P(x',y) \leqslant 1$.

The lower and upper bounds in Equation (5.2) follow directly from the Tian-Pearl bounds on P(benefit). There are two additional possible lower bounds on P(benefit):

- $P(y_x) - P(y)$ and
- $P(y) - P(y_{x'})$.

Notice that,

$$P(y_x) - P(y) = \text{ATE} + P(y_{x'}) - P(y). \tag{5.3}$$

Since it is possible that $P(y_{x'}) - P(y) \geqslant 0$, it might appear that we should add the right side of the equality in Equation (5.3) to the max function of the lower bound in equation (5.2). However, remember that we do not have access to any causal effects, including $P(y_{x'})$, only the ATE and observational data. The causal effect $P(y_{x'})$ could simply be removed and a lower bound of ATE $- P(y)$ could be added to the max function. However, the

existing lower bound ATE subsumes it because $P(y) \geqslant 0$. Similarly, since $P(y) - P(y_{x'}) = \text{ATE} - P(y_x) + P(y)$ and $P(y_x) \leqslant 1$, a potential lower bound of $\text{ATE} - P(y')$ could be added to inequalities (5.2). Again, the existing lower bound ATE subsumes it. Since it is possible that $P(y_{x'}) - P(y) \geqslant 0$, it might appear that we should add the right side of the equality in Equation (5.3) to the max function of the lower bound in equation (5.2). However, remember that we do not have access to any causal effects, including $P(y_{x'})$, only the ATE and observational data. A potential lower bound of $\text{ATE} - P(y)$ could not be added to the max function of the left inequality (5.2). However, the existing lower bound ATE subsumes it because $P(y) \geqslant 0$. Similarly, since $P(y) - P(y_{x'}) = \text{ATE} - P(y_x) + P(y)$ and $P(y_x) \leqslant 1$, a potential lower bound of $\text{ATE} - P(y')$ could be added to inequalities (5.2). Again, the existing lower bound ATE subsumes it. In the following section, we discuss some of the ramifications of these bounds.

There are two additional possible upper bounds on $P(\text{benefit})$ as well:

- $P(y_x)$ and
- $P(y'_{x'})$.

Since $P(y_x) = \text{ATE} + P(y_{x'})$ and $P(y_{x'}) \leqslant 1$, a potential upper bound of $\text{ATE} + 1$ could be added to the min function of the right inequality (5.2). However, the existing upper bound $\text{ATE} + P(x, y') + P(x', y)$ subsumes it because $P(x, y') + P(x', y) \leqslant 1$. Similarly, since $P(y'_{x'}) = \text{ATE} + 1 - P(y_x) = \text{ATE} + P(y'_x)$ and $P(y'_x) \leqslant 1$, a potential upper bound of $\text{ATE}+1$ could be added to inequalities (5.2). Again, the existing upper bound $\text{ATE} + P(x, y') + P(x', y)$ subsumes it.

It may appear that the upper bound might dip below the lower bound, which would be problematic. In particular, either of the following two cases would cause this situation:

$$P(x, y) + P(x', y') < \text{ATE}, \text{or} \tag{5.4}$$

$$\text{ATE} + P(x, y') + P(x', y) < 0. \tag{5.5}$$

Neither of these inequalities can occur because of the inequalities in (5.10). Inequality (5.4) cannot occur because of the right inequality of (5.10). Similarly, inequality (5.5) cannot occur because of the left inequality of (5.10).

5.3 How Observational Data Inform P(benefit)

The bounds on P(benefit) produced by Eqs. (5.1) and (5.2) can be visualized interactively at `https://ate-bounds.streamlit.app` to develop an intuitive feel for these bounds. The graphs in this chapter are taken directly from this visualization tool I created.

To show the contrast between Eq. (5.1) and Eq. (5.2), Fig. 5.1 displays the allowable values of P(benefit) for various levels of ATE, assuming no observational information is available (i.e., Eq. (5.1)). We see, for example, that for ATE $= 0$ (left vertical dashed line), the bound is vacuous ($0 \leqslant P(\text{benefit}) \leqslant 1$), while for ATE $= 0.5$ (right vertical dashed line), we have $\frac{1}{2} \leqslant P(\text{benefit}) \leqslant 1$ — a somewhat more informative bound, but still rather trivial.

Figure 5.2 displays the allowable values of P(benefit) when observational data are available. We see that for ATE $= 0$ (left vertical dashed line), we now have $0 \leqslant P(\text{benefit}) \leqslant \frac{1}{2}$, whereas for ATE $= 0.5$ (right vertical dashed line), we now have a point estimate $P(\text{benefit}) = \frac{1}{2}$, assuring us that exactly 50% of all subjects will benefit from the treatment (and none will be harmed by it).

When observational data become less symmetric, say $P(x) = 0.5$, $P(y|x) = 0.9$, and $P(y|x') = 0.1$, the regions of possible and impossible P(benefit) values shift significantly. Moving the sliders for $P(x)$, $P(y|x)$, and $P(y|x')$ to the above values produces the graph shown in Figure 5.3. This time, when ATE $= 0$ (left vertical dashed line), the bounds on P(benefit) narrow down to $0 \leqslant P(\text{benefit}) \leqslant 0.1$, telling us that subjects have a maximum 10% chance of benefiting from the treatment. When ATE $= 0.5$ (right vertical dashed line), we have $\frac{1}{2} \leqslant P(\text{benefit}) \leqslant 0.6$, still a narrow width of 0.1, with an assurance of at least 50% chance of benefiting from the treatment.

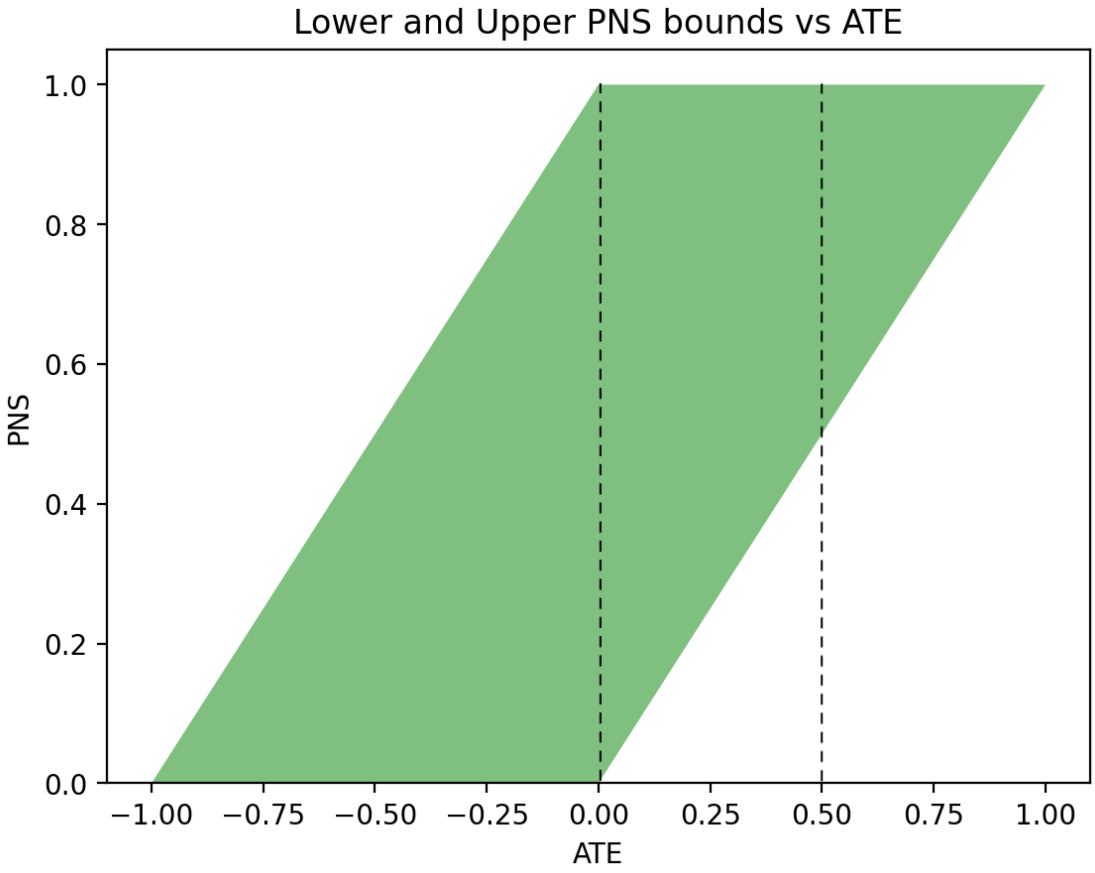


Figure 5.1: The green area represents possible P(benefit) values for the given ATE, while the white areas represent values not achievable by P(benefit).

It should be clear now that consequential information on individual benefit can be obtained even when separate causal effects, $P(y_x)$ and $P(y_{x'})$, are unavailable. The same situation holds for P(harm) as well.

5.4 How Observational Data Inform the Probability of P(harm)

The probability of harm is the converse of P(benefit). We can bound this probability with ATE and with observational data similar to Eqs. (5.1) and (5.2). With just the ATE,

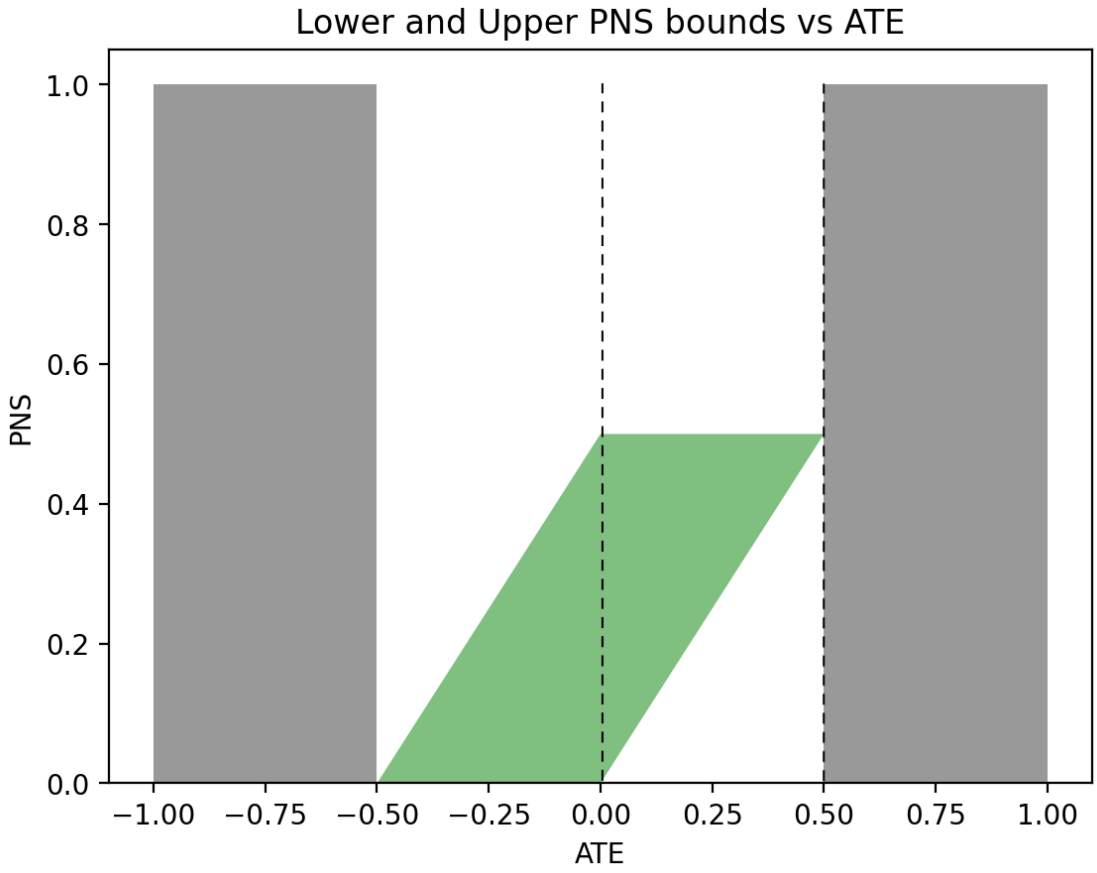


Figure 5.2: The green area represents possible $P(\text{benefit})$ values for the given ATE, while the gray areas represent values of ATE that are incompatible with the assumed observational information: $P(y|x) = P(y|x') = P(x) = 0.5$.

$P(\text{harm})$ is bounded as:

$$\max\{0, -\text{ATE}\} \leqslant P(\text{harm}) \leqslant \min\{1, 1 - \text{ATE}\}. \tag{5.6}$$

The lower bound is positive when ATE is positive, and the upper bound is less than 1 when ATE is negative. When we combine observational data, a smaller upper bound is possible:

$$\max\{0, \text{ATE}\} \leqslant P(\text{harm}) \leqslant \min \left\{ \begin{array}{r} P(x, y') + P(x', y), \\ P(x, y) + P(x', y') - \text{ATE} \end{array} \right\}. \tag{5.7}$$

Again, the upper bound in (5.7), is always lower than (or equal to) the one in (5.6), since

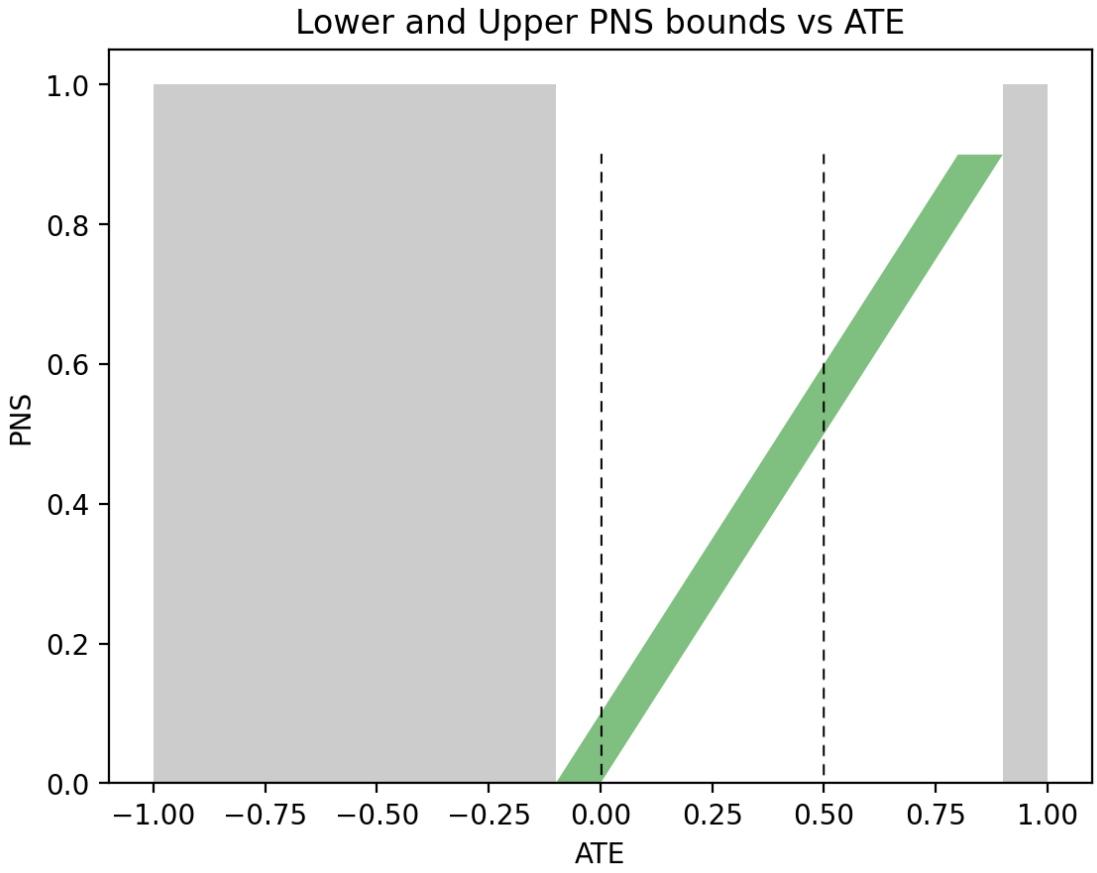
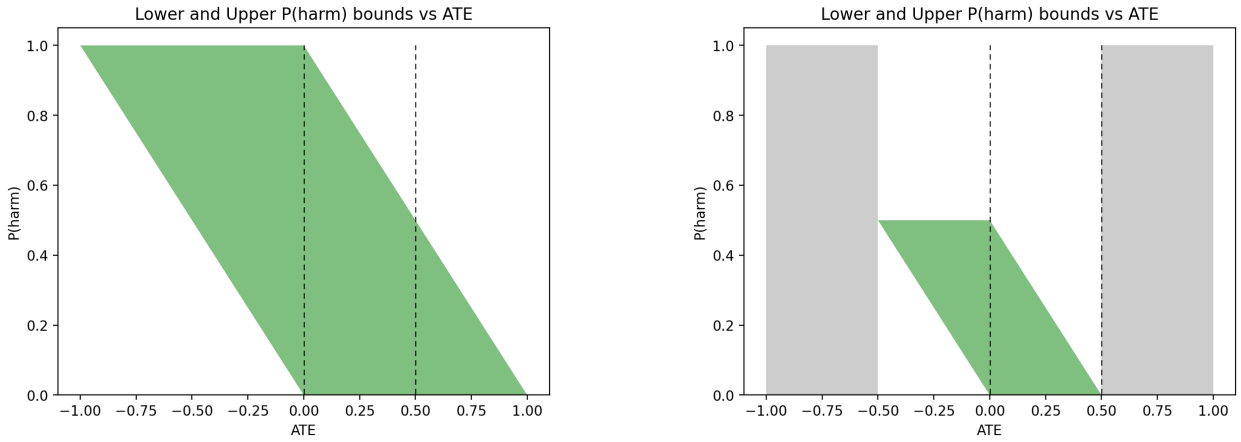


Figure 5.3: The green area represents possible P(benefit) values for the given ATE and observational probabilities: $P(x) = 0.5, P(y|x) = 0.9, P(y|x') = 0.1$.

$P(x, y) + P(x', y') \leqslant 1$ and $P(x, y') + P(x', y) \leqslant 1$. See Appendix 1 for the derivations of Eqs. (5.6) and (5.7).

Figures 5.4a and 5.4b depict these bounds under the same conditions as Figures 5.1 and 5.2, respectively.

Figure 5.5, on the other hand, shows very different sets of bounds for the asymmetric case: $P(x) = 0.5$, $P(y|x) = 0.9$, and $P(y|x') = 0.1$, used in Figure 5.3. This time, $0 \leqslant P$(harm) $\leqslant 0.1$ when ATE $= 0$ or ATE $= 0.5$. This has the same width of 0.1 as in the case of P(benefit), but with a substantially different shape.



(a) No observational data

(b) $P(x) = P(y|x) = P(y|x') = 0.5$

Figure 5.4: P(harm) graphs corresponding to Figures 5.1 and 5.2 for P(benefit)

5.5 Probability of Immunity and Doom Bounds

As learned in Chapter 3, P(immunity) is a simple function of $P(y_x)$ and P(benefit) in Equation (3.24) and P(doom) is a simple function of $P(y_{x'})$ and P(harm) in Equation (3.25). Unfortunately, since the causal effects $P(y_x)$ and $P(y_{x'})$ are unknown, this leaves no room for ATE to play a role in narrowing bounds on P(immunity) and P(doom). We are left with their observational-data-only bounds:

$$0 \leqslant P(\text{immunity}) \leqslant P(y) + P(x, y) + P(x', y'),$$
$$0 \leqslant P(\text{doom}) \leqslant P(y') + P(x, y) + P(x', y').$$

5.6 Intuition

The intuition behind the P(benefit) lower bound is that the probability of benefiting cannot be less than the positive part of the difference in causal effects. That difference must be explained by benefiting from treatment.

The intuition behind the P(benefit) upper bound is split into two parts. First, the

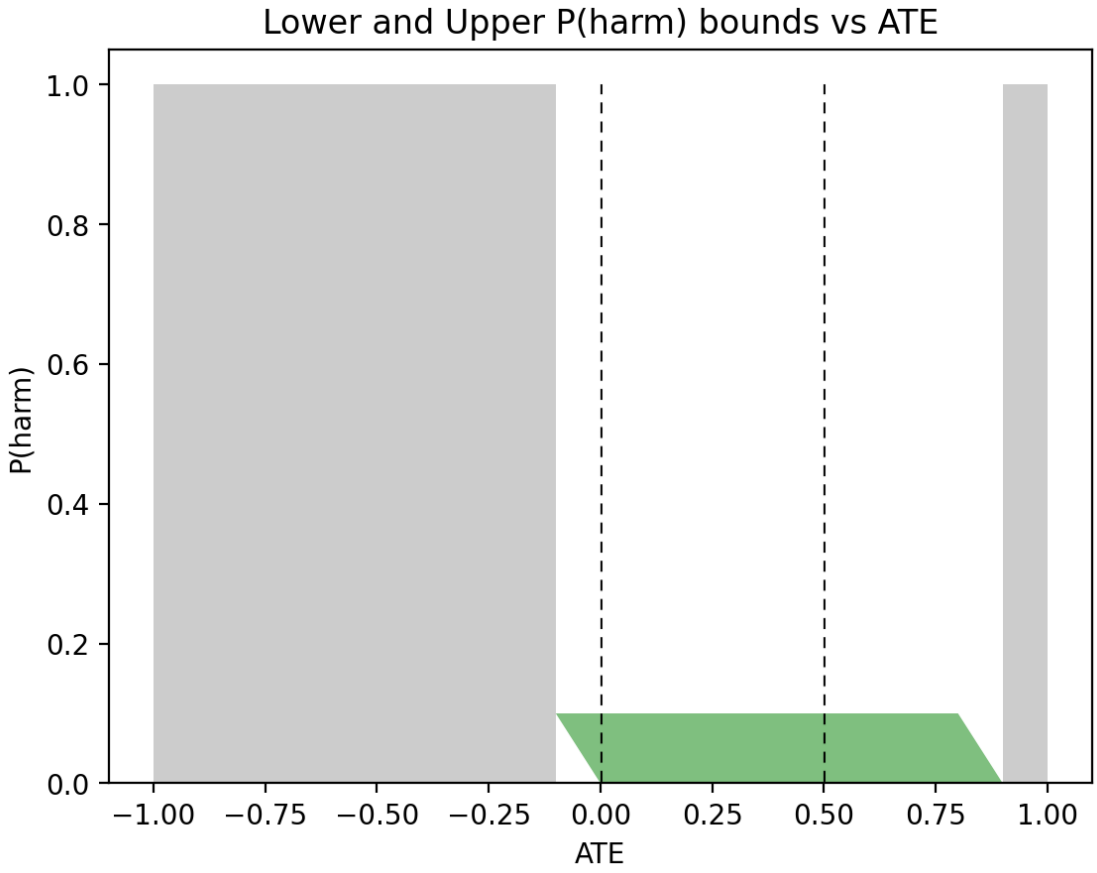


Figure 5.5: The green area represents possible $P(\text{harm})$ values for the given ATE and observational probabilities: $P(x) = 0.5, P(y|x) = 0.9, P(y|x') = 0.1$.

benefiters must be among the individuals who chose treatment and had a successful outcome, (x, y), or those who avoided treatment and had an unsuccessful outcome, (x', y'). Therefore, one potential upper bound is $P(x, y) + P(x', y')$. Alternatively, since ATE $= P(\text{benefit}) - P(\text{harm})$, we get an upper bound on $P(\text{benefit})$ by adding at least the proportion of individuals harmed to ATE. This is precisely the upper bound ATE$+P(x, y')+P(x', y)$ because the proportion of individuals choosing treatment and having an unsuccessful outcome, $P(x, y')$, and the proportion of individuals avoiding treatment and having a successful outcome, $P(x', y)$, comprise all individuals harmed by treatment as well as some additional individuals.

Similar reasoning holds for the lower and upper bounds of P(harm), P(immunity), and P(doom).

5.7 Bounds on ATE

In addition to informing P(benefit) and P(harm), observational data also impose restrictions on ATE, violations of which imply experimental imperfections. We start with Tian and Pearl's bounds on causal effects [TP00]:

$$P(x, y) \leqslant P(y_x) \leqslant 1 - P(x, y'), \tag{5.8}$$

$$P(x', y) \leqslant P(y_{x'}) \leqslant 1 - P(x', y'). \tag{5.9}$$

If we multiply Equation (5.9) by -1 and add it to Equation (5.8), we get the following bounds on ATE:

$$P(x, y) + P(x', y') - 1 \leqslant \text{ATE} \leqslant P(x, y) + P(x', y'). \tag{5.10}$$

While the range of ATE values has a width of 1, the location of this range can still alert the experimenter to possible incompatibilities between the observational and experimental data.

CHAPTER 6

Intention as Evidence

6.1 Introduction

The practical value that knowledge of individualized probabilities of benefit, harm, immunity, and doom brings has been expressed in prior chapters. More practical use of these probabilities of causation can be obtained if an individual's treatment choice is known. This notion is often captured in the Effect of Treatment on the Treated (ETT):

$$\begin{aligned} \text{ETT} &= E[Y_x - Y_{x'}|x] \\ &= E[Y_x|x] - E[Y_{x'}|x]. \end{aligned}$$

Like the ATE, ETT hides information that could be crucial to decision making.

6.2 Treatment Informs PS

To uncover the vital information that ETT hides, ATE can be decomposed into P(benefit) and P(harm) when X and Y are binary:

$$\text{ATE} = P(\text{benefit}) - P(\text{harm}). \tag{6.1}$$

ETT can similarly be decomposed:

$$\text{ETT} = P(\text{benefit}|x) - P(\text{harm}|x), \tag{6.2}$$

$$\begin{aligned} P(\text{benefit}|x) &= P(y_x, y'_{x'}|x) \\ &= P(y, y'_{x'}|x) \\ &= P(y'_{x'}|x, y) \cdot P(y|x) \\ &= \text{PN} \cdot P(y|x), \end{aligned} \tag{6.3}$$

$$\begin{aligned} P(\text{harm}|x) &= P(y'_x, y_{x'}|x) \\ &= P(y', y_{x'}|x) \\ &= P(y_{x'}|x, y') \cdot P(y'|x) \\ &= \text{PN}_{y'} \cdot P(y'|x) \end{aligned} \tag{6.4}$$

where $\text{PN}_{y'}$ is PN with y' swapped with y.

This allows a decision maker to understand the probability of benefiting or being harmed the moment after a decision for treatment has been made. These probabilities of causation can be particularly useful when a decision has been made for no treatment. The analogous situation to the ATE is referred to as the Effect of Treatment on the Untreated (ETU):

$$\text{ETU} = P(\text{benefit}|x') - P(\text{harm}|x'), \tag{6.5}$$

$$\begin{aligned} P(\text{benefit}|x') &= P(y_x, y'_{x'}|x') \\ &= P(y_x, y'|x') \\ &= P(y_x|x', y') \cdot P(y'|x') \\ &= \text{PS} \cdot P(y'|x'), \end{aligned} \tag{6.6}$$

$$\begin{aligned} P(\text{harm}|x') &= P(y'_x, y_{x'}|x') \\ &= P(y'_x, y|x') \\ &= P(y'_x|x', y) \cdot P(y|x') \\ &= \text{PS}_{y'} \cdot P(y|x') \end{aligned} \tag{6.7}$$

where $\text{PS}_{y'}$ is PS with y' swapped with y.

Without assumptions, such as monotonicity, or knowledge of the underlying structural equations, PN and PS can generally only be bounded according to Tian-Pearl bounds:

$$\max\left\{0, \frac{P(y) - P(y_{x'})}{P(x,y)}\right\} \leqslant \text{PN} \leqslant \min\left\{1, \frac{P(y'_{x'}) - P(x',y')}{P(x,y)}\right\}, \tag{6.8}$$

$$\max\left\{0, \frac{P(y_x) - P(y)}{P(x',y')}\right\} \leqslant \text{PS} \leqslant \min\left\{1, \frac{P(y_x) - P(x,y)}{P(x',y')}\right\}. \tag{6.9}$$

Combining Equations (6.3), (6.4), and (6.8) provides $P(\text{benefit})$ and $P(\text{harm})$ on the treated, and combining Equations (6.6), (6.7), and (6.9) provides $P(\text{benefit})$ and $P(\text{harm})$ on the untreated:

$$\max\left\{0, \frac{P(y) - P(y_{x'})}{P(x)}\right\} \leqslant P(\text{benefit}|x) \leqslant \min\left\{P(y|x), \frac{P(y'_{x'}) - P(x',y')}{P(x)}\right\}, \tag{6.10}$$

$$\max\left\{0, \frac{P(y') - P(y'_{x'})}{P(x)}\right\} \leqslant P(\text{harm}|x) \leqslant \min\left\{P(y'|x), \frac{P(y_{x'}) - P(x',y)}{P(x)}\right\}, \tag{6.11}$$

$$\max\left\{0, \frac{P(y_x) - P(y)}{P(x')}\right\} \leqslant P(\text{benefit}|x') \leqslant \min\left\{P(y'|x'), \frac{P(y_x) - P(x,y)}{P(x')}\right\}, \tag{6.12}$$

$$\max\left\{0, \frac{P(y'_x) - P(y')}{P(x')}\right\} \leqslant P(\text{harm}|x') \leqslant \min\left\{P(y|x'), \frac{P(y'_x) - P(x,y')}{P(x')}\right\}. \tag{6.13}$$

6.3 Example

Consider the following hypothetical scenario. A doctor decides to perform surgery on a patient. Before he does, we can analyze the effectiveness of this treatment with the knowledge that surgery has been decided. From experimental data, we know the probabilities of success (y) for surgery (x) and no surgery (x'):

$$P(y_x) = 0.8,$$

$$P(y_{x'}) = 0.6.$$

Additionally, historical electronic health records show the following observational data:

$$P(x) = 0.4,$$
$$P(y|x) = 0.5,$$
$$P(y|x') = 0.7$$

where X represents the doctor's decision.

Many decision-makers would look at the ATE to make a decision for or against surgery:

$$\begin{aligned}\text{ATE} &= P(y_x) - P(y_{x'}) \\ &= 0.2.\end{aligned}$$

The ATE is positive and, neglecting other costs such as financial, pain, and recovery, doctors would choose surgery.

The ATE$|X$ is known in literature as ETT, though it is rarely applied in medical contexts. In many cases, it is not the probability of recovery or probability of death that counts. The P(harm) and P(benefit) are often the true quantities of interest.

A physician might have conscious and unconscious clues informing their opinion. For example, they might see their patient is sweating, the parents are highly educated, or other indirect thoughts the physician is unaware of. These clues might sway the doctor on the basis of their personal experience, to be taken into account in an undisclosed way. How do we combine the hunches of this doctor with the rigorously researched standard of care to come up with a decision that is better than each of the two in isolation?

We can compute the posteriors $P(\text{benefit}|X)$ and $P(\text{harm}|X)$ by treating intention as evidence. Within the time between making a decision and acting on that decision, the updated beliefs in PoCs might make it clear to reverse a decision.

First, let us compute the standard Tian-Pearl bounds on $P(\text{benefit})$. Equation (6.1)

allows a quick derivation of $P(\text{harm})$:

$$\max\{0, 0.2, 0.18, 0.02\} = \mathbf{0.2} \leqslant P(\text{benefit}) \leqslant \mathbf{0.38} = \min\{0.8, 0.4, 0.38, 0.82\},$$
$$0.2 - 0.2 = \mathbf{0} \leqslant P(\text{harm}) \leqslant \mathbf{0.18} = 0.38 - 0.2.$$

Since our hypothetical doctor has already decided on surgery, this decision can be used to obtain more accurate bounds on $P(\text{benefit})$ and $P(\text{harm})$ with Equations (6.10) and (6.11):

$$\max\left\{0, \frac{0.62 - 0.6}{0.4}\right\} = \mathbf{0.05} \leqslant P(\text{benefit}|x) \leqslant \mathbf{0.5} = \min\left\{0.5, \frac{0.4 - 0.18}{0.4}\right\},$$
$$\max\left\{0, \frac{0.38 - 0.4}{0.4}\right\} = \mathbf{0} \leqslant P(\text{harm}|x) \leqslant \mathbf{0.45} = \min\left\{0.5, \frac{0.6 - 0.42}{0.4}\right\}.$$

Surprisingly, both PoCs above have far looser bounds than when not conditioning on x. The extra information about the doctor's decision for surgery counter-intuitively opens up more possibilities for the probabilities of benefit and harm. One way to think about this is that the decision could be a poor decision or a good decision. A poor decision will lower the probability of benefiting, while a good decision will raise it, thus widening the original bounds.

On the other hand, an interesting result appears if the surgeon decides against surgery. Point estimates for both $P(\text{benefit}|x')$ and $P(\text{harm}|x')$ are computed:

$$\max\left\{0, \frac{0.8 - 0.62}{0.6}\right\} = \mathbf{0.3} \leqslant P(\text{benefit}|x') \leqslant \mathbf{0.3} = \min\left\{0.3, \frac{0.8 - 0.2}{0.6}\right\},$$
$$\max\left\{0, \frac{0.2 - 0.38}{0.6}\right\} = \mathbf{0} \leqslant P(\text{harm}|x') \leqslant \mathbf{0} = \min\left\{0.7, \frac{0.2 - 0.2}{0.6}\right\}.$$

Both point estimates are within the original bounds on $P(\text{benefit})$ and $P(\text{harm})$ before any surgery decisions were made. This makes it clear that the original bounds are an average of when surgery is decided and surgery is decided against:

$$P(\text{benefit}) = P(\text{benefit}|x) \cdot P(x) + P(\text{benefit}|x') \cdot P(x'),$$
$$P(\text{harm}) = P(\text{harm}|x) \cdot P(x) + P(\text{harm}|x') \cdot P(x').$$

6.4 No Treatment as Evidence

If we can make the assumption that a decision not to treat can be revisited later with no consequences to the outcome, then we can treat the no treatment decision as evidence. This may be the case when a decision for no surgery means the patient remains with their disease and all else is equal.

The question now becomes, "if I chose not to treat and nothing improved, will I see improvement if I choose treatment now?" This is precisely the PS: $P(y_x|x', y')$. The bounds on PS has already been determined in Equation (6.9) and throughout Chapter 4 it was narrowed in various ways.

If we were able to use no treatment as evidence in the motivating numerical example of Section 3.3, we could save lives. Notice that 9$ of women chose no treatment and did not recover. The PS for men and women are:

$$P(y_x|x', y', \text{female}) = 1,$$
$$P(y_x|x', y', \text{male}) = 0.$$

Applying treatment to women after their no-recovery saves 9% of women's lives. On the other hand, there is clearly no need to subject men to the expense, pain, and frustration of treatment as they will surely not recover.

6.4.1 Narrowing P(benefit) and PN

There is another benefit to follow-up decisions that are unaffected by their previous decisions. Notice that P(benefit) is a function of PN, PS, and observational probabilities:

$$\begin{aligned} P(\text{benefit}) &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'}, x) + P(y_x, y'_{x'}, x') \\ &= P(y'_{x'}, x, y) + P(y_x, x', y') \\ &= \text{PN} \cdot P(x, y) + \text{PS} \cdot P(x', y'). \end{aligned} \tag{6.14}$$

Not treating and then treating, under the assumption that the latter is not affected by the former, provides empirical data for PS. This also assumes that the decision to treat the second time is not confounded with the outcome.

Computing an empirical point estimate of PS allows for potentially narrower bounds on P(benefit) and PN, according to Equation (6.14).

6.5 Summary

There is always a gap between decision and action. This gap allows critical information to stop a bad decision. We can imagine one day that a wearable device can guide us the moment decisions are made. If we find that section of our brain which commits to decisions, this guidance might even be oblivious to us. It will simply seem like we are all just phenomenal decision makers.

Sometimes we are afforded the opportunity to redo our decisions. A mistaken decision is evidence we can compute posteriors from. Even the poorest decision makers among us can still have great outcomes.

CHAPTER 7

Monotonicity

7.1 Introduction

Many reasoning tasks in healthcare, marketing, and economics are plagued with indeterminacies in the sense that point estimates of some probabilities cannot be obtained even with infinite data. Instead, ranges of values can be derived, but these are often too wide to be useful.

A common thread among these tasks is that indeterminacies are alleviated or eliminated when monotonicity is assumed (i.e. that outputs can never decrease when inputs increase). For example, that no patient can be harmed by a certain treatment, or that no customer will churn only when offered an incentive. A formal definition of this notion will be given in Section 7.2 together with formulas that connect this to the observed data.

To illustrate the role of monotonicity, we first discuss the problem of unit selection [LP19]. Here the goal is to maximize the gain f associated with a set of units (e.g. patients, customers, or voters) each of them may either benefit from, be harmed by, or remain unaffected by an action under consideration (e.g. treatment, advertisement, or policy). The overall gain, $f(\beta, \gamma, \theta, \delta)$, depends on four parameters: the gain of selecting a unit benefiting from treatment (β), the gain of selecting a unit always having a positive outcome regardless of treatment (γ), the gain of selecting a unit always having a negative outcome regardless of treatment (θ), and the gain of selecting a unit harmed by treatment (δ). Li showed that when monotonicity does not hold, the overall gain f cannot be point estimated from experimental

data alone, as practiced in A/B testing. Moreover, A/B testing, which has been the mainstay of marketing, product development, and other business optimizations, may be grossly sub-optimal, leading to regrettable decisions. Fortunately, the assumption of monotonicity renders optimizations based on A/B testing equivalent to optimizing $f(\beta, \gamma, \theta, \delta)$ over its four parameters. Given data from several sources we can identify when this equivalence holds.

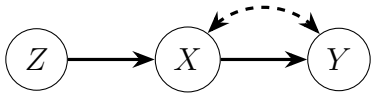


Figure 7.1: Typical structure for IV methods where Z is an instrument for the relationship between X and Y, shown to be marred by unobserved confounders (bidirectional arrow).

A second task demanding the assumption of monotonicity is Instrumental Variable (IV) analysis. IV analysis is also possible if effect homogeneity holds instead of monotonicity. However, [HR20] notes that, "homogeneity is often an implausible condition," whereas, "monotonicity [appears] credible in many settings." The purpose of IV analysis is to estimate the ATE in situations where unobserved confounders exist between treatment X and outcome Y, as shown in Figure 7.1. When monotonocity can be assumed between the instrument, Z, and the treatment variable, X, the ATE can be identified in certain subpopulations, called "compliers". A unit is called a complier if treatment is taken if and only if it is assigned to that unit. This ATE among compliers is referred to as the Local Average Treatment Effect (LATE) and is computed by:

$$\text{LATE} = E[Y_x - Y_{x'}|\text{complier}] = \frac{E[Y|z] - E[Y|z']}{E[X|z] - E[X|z']} \tag{7.1}$$

where Y_x is the value of Y had X been x (treatment) and $Y_{x'}$ is the value of Y had X been x' (non-treatment). Naturally, this is vital in disciplines where confounding is difficult to deal with, such as experimental econometrics [IA94] and social sciences [MW14].

Monotonicity is required to assure the validity of Equation (7.1). Absent monotonity, the denominator of (7.1) may blow up LATE, which further distances LATE from ATE. For

a formal definition of IV and its various extensions using graphs see [Pea09, Chapter 7] and [Pea11].

A third task where monotonicity plays an important role is Causes of Effects (CoE) analysis, which aims to estimate the probability that one event is a "cause" of an observed outcome. This was discussed in Chapter 4. Examples are assigning credit and blame in legal situations, medical diagnosis, and system troubleshooting. These applications invoke counterfactual reasoning and therefore the desired probabilities cannot be determined from either experimental or observational data. Counterfactual probabilities in common use are PN, PS, and PNS (aka P(benefit).

Tian and Pearl [TP00] derived tight bounds on Probabilities of Causation on the basis of experimental and observational data and this dissertation has explored many techniques to further narrow those bounds by, in part, appealing to the causal structure when such is available. These bounds are often still too loose to be useful. If monotonicity can be assumed, however, the bounds collapse to a point [Pea99] based on experimental data alone, even without considering the causal structure. If an identifiable causal structure can additionally be assumed on top of monotonicity then these PoCs are identified with just observational data.

Given its ubiquity in interpreting experimental studies, the need arises to determine when monotonicity is violated, when it can possibly be presumed to hold, and when it definitely holds. In some cases monotonicity is self-evident, for example, in advertising a new product. The control group, not given the information about the product, has no way of purchasing it. Monotonicity must hold because $P(Y_{\text{no ad}} = \text{purchase}) = 0$. In general, however, monotonicity cannot be assured a priori. In medicine, for example, a person might have a 5% chance of being harmed by treatment and a 10% chance of benefiting from it, which may result in a lawsuit if an autopsy proves the former.

This chapter shows how data can be assessed for monotonicity. A *necessary test* indicates when monotonicity is possible and a *sufficiency test* indicates when monotonicity is assured

(Section 7.2). An accompanying interactive plot visualizes how necessity and sufficiency depend on experimental and observational data available (Section 7.3).

7.2 Monotonicity Tests, Sufficiency and Necessity

Let us denote the variables $X \in \{x, x'\}$ and $Y \in \{y, y'\}$ as binary treatment and recovery, respectively. The values x and x' may represent treatment and no treatment, and the values y and y' may represent recovery and no recovery. I will further use y_x to denote the counterfactual sentence, "Variable Y would have the value y, had X been x." Extensions to multi-valued ordinal outcomes will be discussed in Chapter 9.

Using these binary variables, in addition to Definition 4.2.4, monotonicity is defined as,

$$P(y'_x, y_{x'}) \stackrel{\text{def}}{=} P(\text{harm}) = 0. \tag{7.2}$$

A properly conducted RCT yields unbiased estimates of $P(y_x)$ and $P(y_{x'})$, from which we can obtain the ATE. In contrast, observational studies provide estimates of the joint distribution $P(X, Y)$, from which we can obtain $P(x)$, $P(y)$, $P(y|x)$, and $P(y|x')$. As discussed in Chapter 3 an RCT does not directly inform us about $P(\text{harm})$, nor about the other three response types:

$$P(\text{benefit}) \stackrel{\text{def}}{=} P(y_x, y'_{x'}), \tag{7.3}$$

$$P(\text{immunity}) \stackrel{\text{def}}{=} P(y_x, y_{x'}), \tag{7.4}$$

$$P(\text{doom}) \stackrel{\text{def}}{=} P(y'_x, y'_{x'}). \tag{7.5}$$

As a consequence, in contrary to a prevailing myth, ATE does not represent the proportion of people benefiting from treatment. Note also that the four probabilities above must sum to 1 and that ATE is related to $P(\text{harm})$ and $P(\text{benefit})$ (this derivation was shown in Equation (3.22)) via:

$$P(\text{harm}) = P(\text{benefit}) - \text{ATE}. \tag{7.6}$$

Equation (7.6) tells us immediately that under monotonicity, P(benefit) coincides with ATE, or, in other words, ATE constitutes a point estimate of P(benefit). More generally, it allows us to compute P(harm) from P(benefit) and ATE, which we will use to define the level of monotonicity violation.

Given these definitions, the question of whether monotonicity is testable can be answered by examining the bounds on P(harm) and asking what conditions would guarantee an upper bound of 0. Given both observational and experimental studies, the bounds on the probability of harm, derived in Equation (3.23), are:

$$\max \left\{ \begin{array}{c} 0, \\ P(y_{x'}) - P(y_x), \\ P(y) - P(y_x), \\ P(y_{x'}) - P(y) \end{array} \right\} \leqslant P(\text{harm}) \leqslant \min \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x, y') + P(x', y), \\ P(y_{x'}) - P(y_x) + \\ P(x, y) + P(x', y') \end{array} \right\}. \tag{7.7}$$

We see that, when $P(y_x) \geqslant P(y_{x'})$ (or ATE is non-negative), a sufficient condition for monotonicity to hold is that at least one of the arguments to the min function be 0. We can summarize this in a theorem.

Theorem 7.2.1 (Monotonicity Sufficiency Test)**.** *Y is monotonic relative to X if*

$$P(y_{x'}) = 0, \ \textit{or} \tag{7.8}$$

$$P(y_x) = 1, \ \textit{or} \tag{7.9}$$

$$P(x, y') = P(x', y) = 0, \ \textit{or} \tag{7.10}$$

$$P(y_x) - P(y_{x'}) = P(x, y) + P(x', y'). \tag{7.11}$$

Note that the left side of Equation (7.11) is the ATE. When $P(y_x) < P(y_{x'})$ (ATE is negative), monotonicity must fail because Equation (7.6) shows that P(harm) must turn positive.

Unfortunately, conditions (7.8), (7.9), (7.10), and (7.11) are in the form of equalities and therefore can only materialize in rare cases. In contrast, lack of monotonicity is easier to verify. For this purpose I devise a necessary test for monotonicity, which identifies the requirements for monotonicity to be possible. This test is more informative and is derived by checking if all arguments to the max function in the lower bound of $P(\text{harm})$ are non-positive:

$$\begin{aligned} P(y_{x'}) &\leqslant P(y_x), \text{ and} \\ P(y) &\leqslant P(y_x), \text{ and} \\ P(y_{x'}) &\leqslant P(y). \end{aligned}$$

This can be put into a more succinct form [Pea09, p. 294], as shown in Theorem 7.2.2.

Theorem 7.2.2. *(Monotonicity Necessity Test) Y is monotonic relative to X only if*

$$P(y_x) \geqslant P(y) \geqslant P(y_{x'}). \tag{7.12}$$

This is useful for two reasons. First, it can quickly eliminate the possibility of monotonicity by checking for three simple parameters in the data. Second, non-monotonicity implies the existence of subpopulations whose reaction to treatment is substantially different, which, in turn, informs us where the mechanism responsible for that variability could be.

7.3 Interactive Plot

At the webpage `https://lbmaps.web.app/mns.html`, I provide an interactive plot that visualizes regions of data which are necessary or sufficient for monotonicity, as we navigate the terrain of experimental and observational data available.

Figure 7.2 provides a snapshot of this interactive plot. The white regions represent conditions that are required for monotonicity to hold. In other words, finding data outside

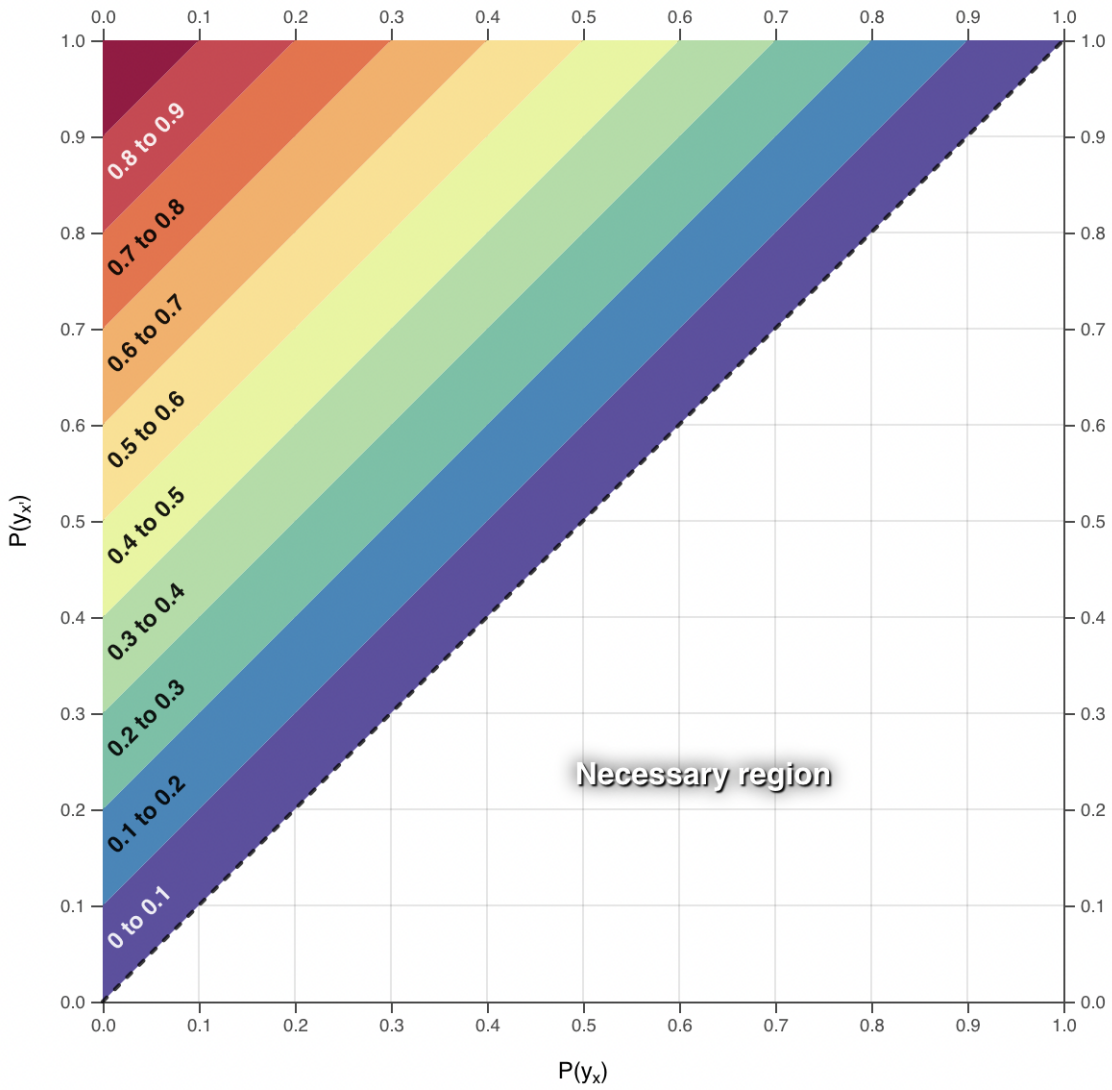


Figure 7.2: Assuming no observational data, it is necessary for $(P(y_x), P(y_{x'}))$ to be in the white region for monotonicity to hold. The color bands represent the minimum degree to which monotonicity is violated for each $(P(y_x), P(y_{x'}))$ combination.

this region implies the existence of units which can be harmed by treatment. Figure 7.2 shows this "necessary" region which, in the absence of observational data, is lying below the dashed diagonal line ($P(y_x) \geqslant P(y_{x'})$). Colored bands indicate data regions where monotonicity definitely does not hold and the color in each band indicates the minimum fraction of violation realizable in that band.

Figure 7.3 shows how the "necessary region" changes when observational data are added. For example, having obtained the additional information of $P(x) = P(y|x) = P(y|x') = 0.5$, the necessary region shrinks to $0.5 \leqslant P(y_x) \leqslant 0.75$ and $0.25 \leqslant P(y_{x'}) \leqslant 0.5$. The transparent gray region indicates areas where $P(y_x)$ and $P(y_{x'})$ are incompatible with the observational data. This could happen, for example, when the population recruited for the experiments is totally different than the one used in the observational study, perhaps due to selection bias. Techniques for detection (see Chapter 8) and overcoming selection bias, under certain conditions, are reported in [BTP14].

In comparison, to identify the vanishingly small regions of sufficiency (where monotonicity must hold), we have to look at the lines marked in red in Figure 7.4. The rarity of this condition is clear since the region is confined to the edges of the purple band ($P(y_{x'}) = 0$ or $P(y_x) = 1$). With observational data of $P(x) = P(y|x) = P(y|x') = 0.5$, this region shrinks to a single point at the bottom right of the compatible region, as seen in Figure 7.5.

7.4 Utilizing Causal Models

We can detect and refute monotonicity with versions of the Monotonicity Sufficiency Test (MST) and Monotonicity Necessity Test (MNT) that take into account a causal model.

7.4.1 Monotonicity Sufficiency Test with Causal Model

One way to utilize a causal model to improve the Monotonicity Sufficiency Test involves decomposing the upper bound of P(harm) in a manner similar to how P(benefit) was de-

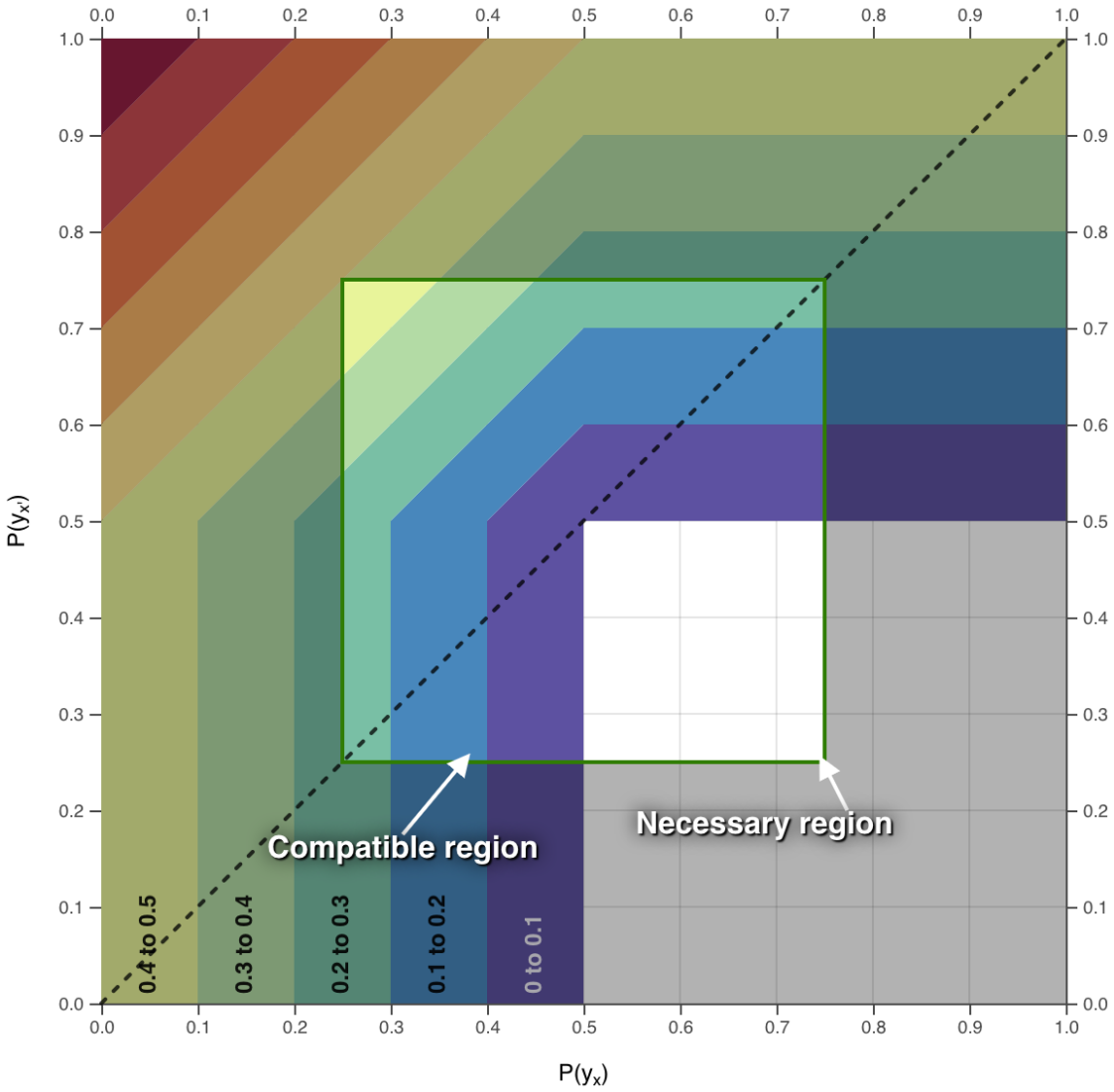


Figure 7.3: Chart showing the impact of observational data on minimum probability of harm. The square in the middle, labeled "Compatible region", indicates values of $P(y_x)$ and $P(y_{x'})$ which are compatible with the observational data $P(x) = P(y|x) = P(y|x') = 0.5$. Incompatibility implies experimental imperfections. The white square, labeled "Necessary region", indicates where monotonicity may hold. The colors in each band indicate the minimum probability of harm (Equation (3.19)).

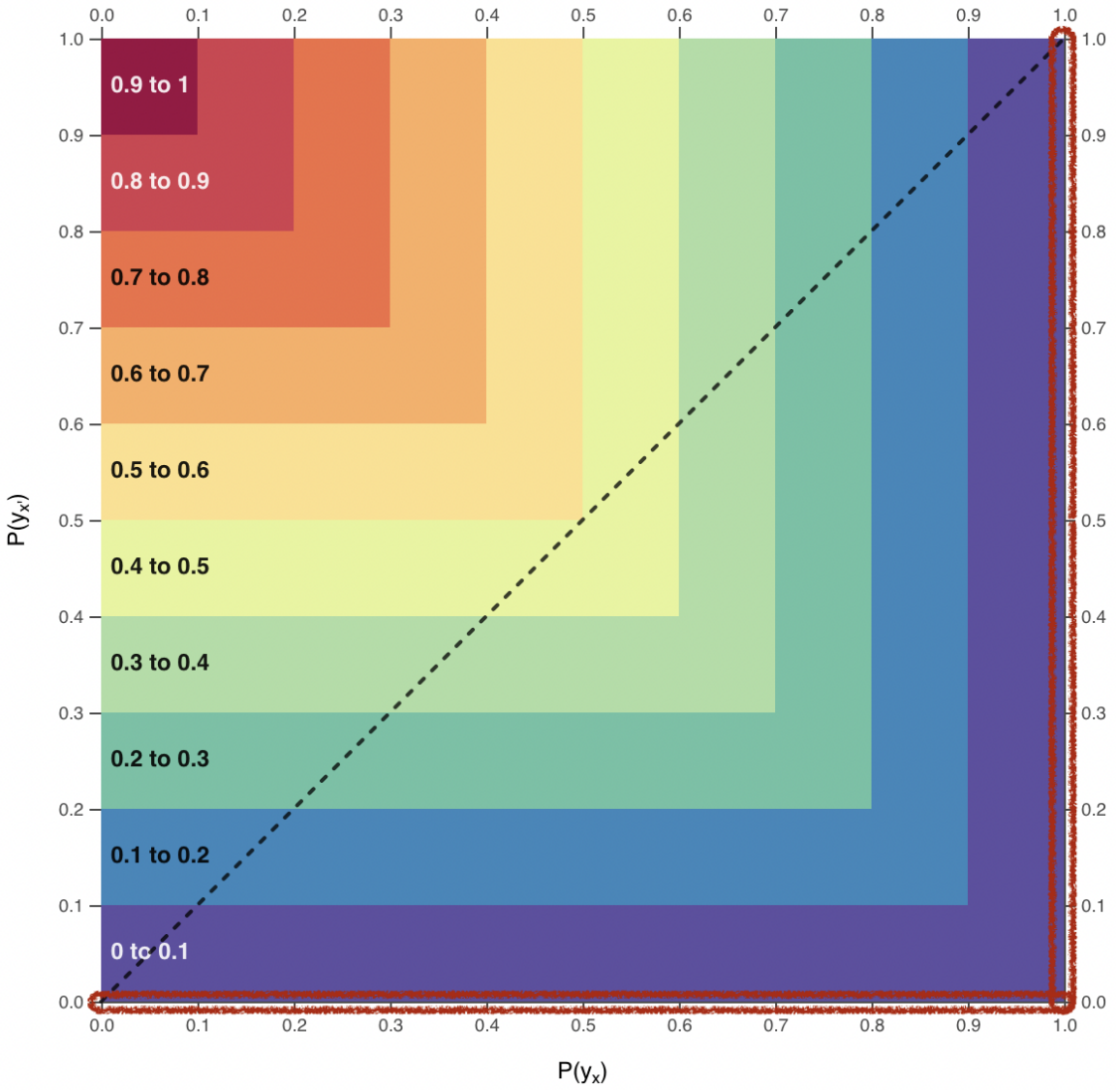


Figure 7.4: Chart showing maximum probability of harm with no observational data. To guarantee monotonicity ($P(\text{harm}) = 0$), $(P(y_x), P(y_{x'}))$ must be on the bottom or right edge of the chart (outlined in red).

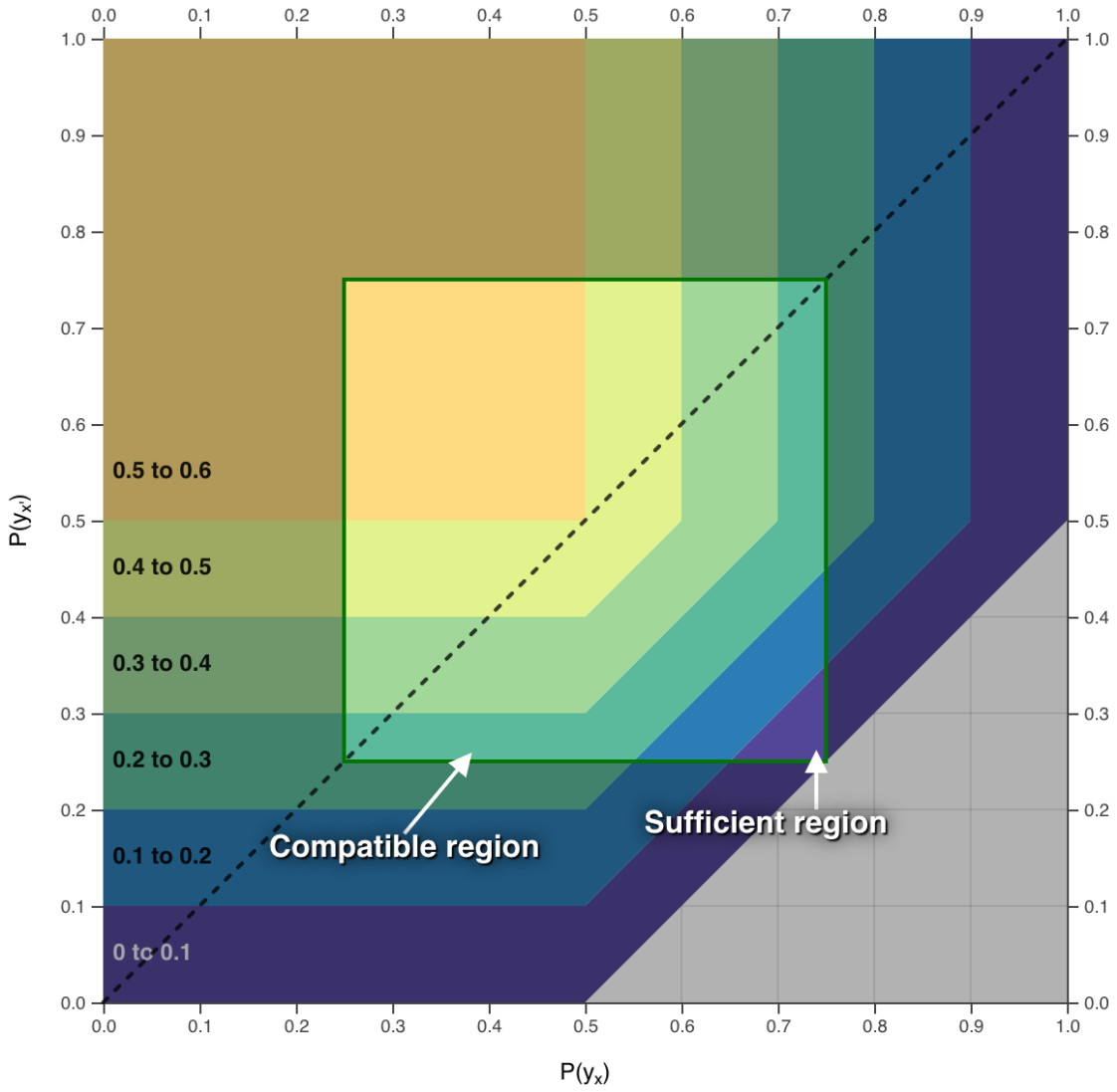


Figure 7.5: Chart showing the impact of observational data on maximum probability of harm. $(P(y_x), P(y_{x'}))$ is only possible in the center square region if $P(x) = P(y|x) = P(y|x') = 0.5$ and it is sufficient for monotonicity at only one point, the bottom right corner of this compatible region.

composed in Theorem 4 of [MLP22].

Lemma 7.4.1. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables not including X and Y, then $P(harm)$'s upper bound is as follows:*

$$P(harm) \leqslant \sum_{\mathbf{z}} \min \left\{ \begin{array}{c} P(y_{x'}|\mathbf{z}), \\ P(y'_x|\mathbf{z}), \\ P(x, y'|\mathbf{z}) + P(x', y|\mathbf{z}), \\ P(y_{x'}|\mathbf{z}) - P(y_x|\mathbf{z}) \ + P(x, y|\mathbf{z}) + P(x', y'|\mathbf{z}) \end{array} \right\} \times P(\mathbf{z}). \tag{7.13}$$

Lemma 7.4.1 allows us to check whether $P(harm)$ is definitely 0 by finding a single argument to the min function in (7.13) that equals 0 for every instantiation of $\mathbf{Z}$. Theorem 7.2.1 can now be enhanced to take into account a causal model and detect monotonicity more often.

Theorem 7.4.1. *(Monotonicity Sufficiency Test with Covariates) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables not including X and Y, then Y is monotonic relative to X if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(\mathbf{z}) = 0, \ \textit{or}$$

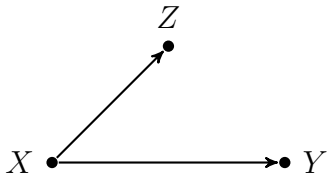
$$P(y_{x'}|\mathbf{z}) = 0, \ \textit{or} \tag{7.14}$$

$$P(y_x|\mathbf{z}) = 1, \ \textit{or} \tag{7.15}$$

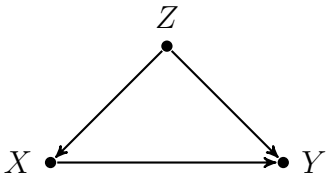
$$P(x, y'|\mathbf{z}) = P(x', y|\mathbf{z}) = 0, \ \textit{or} \tag{7.16}$$

$$P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}) = P(x, y|\mathbf{z}) + P(x', y'|\mathbf{z}). \tag{7.17}$$

Theorem 7.4.1 supersedes Theorem 7.2.1 because they are equivalent when $\mathbf{Z} = \emptyset$. As mentioned in [MLP22], a significant qualification to the causal model needs to be addressed. $P(Y_X|\mathbf{Z})$ will often be unmeasurable if $\mathbf{Z}$ contains descendants of X. This is because if X



(a) $Y_x \perp\!\!\!\perp \mathbf{Z}$ and $\mathbf{Z}$ are descendants of X



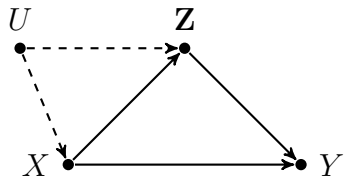
(b) $\mathbf{Z}$ satisfies the back-door criterion

Figure 7.6: Covariate $\mathbf{Z}$ as a descendant of X or confounder of X and Y

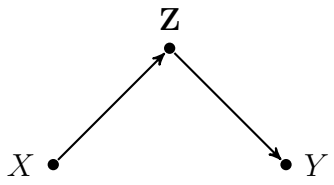
was set to x and $\mathbf{Z}$ contains a descendant of X, then $\mathbf{Z}$ could be altered as well. In this case, $P(Y_X|\mathbf{Z})$ would be a counterfactual query, with a conflict between the observed value of $\mathbf{Z}$ and the value of $\mathbf{Z}$ had X been set to x. A more explicit probability equivalent to $P(y_x|\mathbf{z})$ is $P(y_{x,\mathbf{Z}_x}|\mathbf{z})$, where $\mathbf{Z}_x$ makes $P(y_{x,\mathbf{Z}_x}|\mathbf{z})$ the probability of $Y = y$ had X been set to x and $\mathbf{Z}$ was set to its natural value after setting X to x, conditioned on $\mathbf{Z} = \mathbf{z}$. Therefore, if $\mathbf{Z}$ contains descendants of X and $P(Y_X|\mathbf{Z})$ cannot be measured because of those descendants, then those descendants should be removed from $\mathbf{Z}$.

One scenario that allows a causal Bayesian network to infer $P(y_x|\mathbf{z})$ when $\mathbf{Z}$ contains descendants of X is when the conditional probability tables (CPTs) for $\mathbf{Z}$ and their ancestors are deterministic (probabilities are all 0 or 1). If all endogenous variables are deterministic and marginal probabilities of exogenous variables are known, then the original counterfactual probability is estimable. Another scenario where $P(y_x|\mathbf{z})$ is estimable is when $P(\mathbf{z}_x) = 1$. Lastly, if the descendants of X in $\mathbf{Z}$ are independent of Y_x, such as in Figure 7.6a, then $P(y_x|\mathbf{z})$ would be measurable. However, those descendants would not contribute to any narrowing of bounds. To see this, recognize that $P(y_x|\mathbf{z}^*) = P(y_x)$ if $\mathbf{Z}^* \perp\!\!\!\perp Y_x$, where $\mathbf{Z}^*$ is the set of descendants of X in $\mathbf{Z}$.

If $\mathbf{Z}$ satisfies the back-door criterion [Pea09], such as in Figure 7.6b, then observational data alone is sufficient to potentially detect monotonicity. Similar to how P(benefit) was decomposed with a set of variables satisfying the back-door criterion in Theorem 5 of [MLP22], we can decompose the upper bound of P(harm).



(a) $X \to Y$ and no $\mathbf{Z} - Y$ confounding



(b) $X \not\to Y$ and no confounding

Figure 7.7: Mediator $\mathbf{Z}$

Lemma 7.4.2. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables satisfying the back-door criterion, then $P(harm)$'s upper bound is as follows:*

$$P(harm) \leqslant \sum_{\mathbf{z}} \min\{P(y|x', \mathbf{z}), P(y'|x, \mathbf{z})\} \times P(\mathbf{z}). \tag{7.18}$$

This leads directly to the next theorem.

Theorem 7.4.2. *(Monotonicity Sufficiency Test with Covariates Satisfying Back-door Criterion) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables satisfying the back-door criterion, then Y is monotonic relative to X if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(\mathbf{z}) = 0, \ \textit{or}$$

$$P(y|x', \mathbf{z}) = 0, \ \textit{or} \tag{7.19}$$

$$P(y|x, \mathbf{z}) = 1. \tag{7.20}$$

7.4.1.1 Partial Mediators

Due to the qualification noted above for Theorem 7.4.1 involving descendants of X in $\mathbf{Z}$, along with the exclusion of descendants of X in $\mathbf{Z}$ for Theorem 7.4.2, the MSTs in Theorems 7.4.1 and 7.4.2 are often not applicable. Instead, we can improve on the MST in Theorem 7.2.1 when $\mathbf{Z}$ consists of mediators of X and Y.

In the absence of confounders between $\mathbf{Z}$ and Y and between X and Y, such as in Figure 7.7a, we can bound P(benefit) as follows:

Lemma 7.4.3. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup \mathbf{Z}_{x'} \mid \mathbf{Z}_x)$ in G, then $P(harm)$ is upper bounded as follows:*

$$P(harm) \leqslant \min \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x, y') + P(x', y), \\ P(y_{x'}) - P(y_x) + P(x, y) + P(x', y'), \\ \sum_{\mathbf{z}} \sum_{\mathbf{z}'} \min\{P(y|x', \mathbf{z}), P(y'|x, \mathbf{z}')\} \times \\ \min\{P(\mathbf{z}_{x'}), P(\mathbf{z}'_x)\} \end{array} \right\}. \tag{7.21}$$

This leads directly to the next theorem.

Theorem 7.4.3. *(Monotonicity Sufficiency Test with Partial Mediators) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup \mathbf{Z}_{x'} \mid \mathbf{Z}_x)$ in G. Then Y is monotonic relative to X if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(y|x, \mathbf{z}') = 1, \text{ or} \tag{7.22}$$

$$P(y|x', \mathbf{z}) = 0, \text{ or} \tag{7.23}$$

$$P(\mathbf{z}_x) = 1, \text{ or} \tag{7.24}$$

$$P(\mathbf{z}_{x'}) = 0. \tag{7.25}$$

7.4.1.2 Complete Mediators

If $\mathbf{Z}$ is a complete mediator, such that X affects Y only through $\mathbf{Z}$, there is no confounding between X and $\mathbf{Z}$, and there is no confounding between $\mathbf{Z}$ and Y, then we can obtain a lower upper bound on P(harm).

Lemma 7.4.4. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables such that $(Y \perp\!\!\!\perp X \mid \mathbf{Z}) \wedge (\forall x \in X : Z_x \perp\!\!\!\perp X) \wedge (\forall \mathbf{z} \in \mathbf{Z} : Y_z \perp\!\!\!\perp Z)$, then $P(harm)$ is upper bounded as follows:*

$$P(harm) \leqslant \min \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x, y') + P(x', y), \\ P(y_{x'}) - P(y_x) + P(x, y) + P(x', y'), \\ \sum_{\mathbf{z}} \sum_{\mathbf{z}' \neq \mathbf{z}} \min\{P(y|\mathbf{z}), P(y'|\mathbf{z}')\} \times \\ \min\{P(\mathbf{z}|x'), P(\mathbf{z}'|x)\} \end{array} \right\}. \tag{7.26}$$

This leads directly to the next theorem.

Theorem 7.4.4. *(Monotonicity Sufficiency Test with Complete Mediators) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup \mathbf{Z}_{x'} \mid \mathbf{Z}_x)$ and completely mediates all direct effects of X on Y in G. Then Y is monotonic relative to X if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(y|\mathbf{z}') = 1, \ \textit{or} \tag{7.27}$$

$$P(y|\mathbf{z}) = 0, \ \textit{or} \tag{7.28}$$

$$P(\mathbf{z}|x) = 1, \ \textit{or} \tag{7.29}$$

$$P(\mathbf{z}|x') = 0, \tag{7.30}$$

where $\mathbf{z}' \neq \mathbf{z}$.

Note that Theorem 7.4.4 only involves observational data.

7.4.2 Monotonicity Necessity Test with Causal Model

The corresponding lower bound to P(harm) for the upper bound in Lemma 7.4.1 is the following.

Lemma 7.4.5. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables not including X and Y, then $P(harm)$'s lower bound is as follows:*

$$P(harm) \geqslant \sum_{\mathbf{z}} \max \left\{ \begin{array}{c} 0, \\ P(y_{x'}|\mathbf{z}) - P(y_x|\mathbf{z}), \\ P(y|\mathbf{z}) - P(y_x|\mathbf{z}), \\ P(y_{x'}|\mathbf{z}) - P(y|\mathbf{z}) \end{array} \right\} \times P(\mathbf{z}). \tag{7.31}$$

Lemma 7.4.5 allows us to check whether $P(\text{harm})$ is possibly 0 by checking whether every argument to the max function in (7.31) is less than or equal to 0 (or $P(\mathbf{z}) = 0$) for every instantiation of $\mathbf{Z}$. Theorem 7.2.2 can now be enhanced to take into account a causal model and detect violations of monotonicity more often.

Theorem 7.4.5. *(Monotonicity Necessary Test with Covariates) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables not including X and Y, then Y is monotonic relative to X only if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(\mathbf{z}) = 0, \ \textit{or}$$
$$P(y_x|\mathbf{z}) \geqslant P(y|\mathbf{z}) \geqslant P(y_{x'}|\mathbf{z}). \tag{7.32}$$

If $\mathbf{Z}$ satisfies the back-door criterion, such as in Figure 7.6b, then observational data alone is sufficient to potentially detect violations of monotonicity. Similar to how $P(\text{benefit})$ was decomposed with a set of variables satisfying the back-door criterion in Theorem 5 of [MLP22], we can decompose the lower bound of $P(\text{harm})$.

Lemma 7.4.6. *Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables satisfying the back-door criterion, then $P(harm)$'s lower bound is as follows:*

$$P(harm) \geqslant \sum_{\mathbf{z}} \max\{0, P(y|x', \mathbf{z}) - P(y|x, \mathbf{z})\} \times P(\mathbf{z}). \tag{7.33}$$

This leads directly to the next theorem.

Theorem 7.4.6. *(Monotonicity Necessity Test with Covariates Satisfying Back-door Criterion) Given a causal diagram G and distribution compatible with G, let $\mathbf{Z}$ be a set of variables satisfying the back-door criterion, then Y is monotonic relative to X only if $\forall \mathbf{z} \in \mathbf{Z}$:*

$$P(\mathbf{z}) = 0, \ \textit{or}$$
$$P(y|x, \mathbf{z}) \geqslant P(y|x', \mathbf{z}). \tag{7.34}$$

7.5 ϵ-Limited Harm

In the event that monotonicity exists between two variables, proving monotonicity from data, even with the luxury of both experimental and observational data, can be difficult, as demonstrated by the edges of the plot in Figure 7.4 above and Theorem 7.2.1. With the additional assumption of causal structure, Theorems 7.4.1, 7.4.2, 7.4.3, and 7.4.4 allow for more opportunities to detect monotonicity, however, detection may still elude us.

Allowing for some units or individuals to be harmed, or only part of the population to be monotonic, may result in a more informed distribution of the beneficiaries, which can be utilized in policy making. Let us call the new bounds induced by such allowance ϵ-bounds. Theorems 7.5.1 and 7.5.2 reflect this less restrictive form:

Theorem 7.5.1. *(Sufficiency Test for ϵ-Limited Harm) If the following conditions hold, then ϵ must be a maximum proportion of units harmed:*

$$P(y_{x'}) \leqslant \epsilon, \ \textit{or} \tag{7.35}$$
$$P(y_x) \geqslant 1 - \epsilon, \ \textit{or} \tag{7.36}$$
$$P(x, y') + P(x', y) \leqslant \epsilon, \ \textit{or} \tag{7.37}$$
$$P(y_x) - P(y_{x'}) \geqslant P(x, y) + P(x', y') - \epsilon. \tag{7.38}$$

Theorem 7.5.2. *(Necessary Test for ϵ-Limited Harm) If ϵ is a maximum proportion of units harmed, then the following conditions must hold:*

$$\begin{aligned} P(y_{x'}) &\leqslant P(y_x) + \epsilon, \ \textit{and} \\ P(y) &\leqslant P(y_x) + \epsilon, \ \textit{and} \\ P(y_{x'}) &\leqslant P(y) + \epsilon. \end{aligned}$$

The proof of Theorems 7.5.1 and 7.5.2 follow directly from the bounds expressed in Equation (3.23).

7.5.1 ϵ-Bounds on P(benefit)

Assumptions of ϵ-limited harm can be used to narrow the bounds on P(benefit). From Equation (7.6), P(benefit) can be expressed in terms of ATE and P(harm):

$$P(\text{benefit}) = \text{ATE} + P(\text{harm}). \tag{7.39}$$

Assuming the inequality $0 \leqslant P(\text{harm}) \leqslant \epsilon$ gives:

$$\begin{aligned} \text{ATE} \leqslant P(\text{benefit}) &\leqslant \text{ATE} + \epsilon, \\ P(y_x) - P(y_{x'}) \leqslant P(\text{benefit}) &\leqslant P(y_x) - P(y_{x'}) + \epsilon. \end{aligned} \tag{7.40}$$

Tian and Pearl's tight bounds on PNS [TP00] resemble those in Equation (3.23), with x and x' swapped, and are expressed as:

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y_x) - P(y), \\ P(y) - P(y_{x'}) \end{array} \right\} \leqslant P(\text{benefit}) \leqslant \min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + \\ \quad P(x,y') + P(x',y) \end{array} \right\}. \tag{7.41}$$

The lower bound of P(benefit) in (7.41) already includes ATE as an argument to its max function, so Equation (7.40) cannot help with the lower bound. However, Equation (7.40) can potentially lower the upper bound of P(benefit) in (7.41) by adding the right side of (7.40) as an argument to the min function in (7.41):

$$\max\left\{\begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y_x) - P(y), \\ P(y) - P(y_{x'}) \end{array}\right\} \leqslant P(\text{benefit}) \leqslant \min\left\{\begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + \\ \quad P(x,y') + P(x',y), \\ P(y_x) - P(y_{x'}) + \epsilon \end{array}\right\}. \tag{7.42}$$

Note that any reduced bounds on P(benefit) due to ϵ-limited harm must come from domain-specific knowledge outside the experimental and observational data. The ϵ obtained from data will merely replicate Tian-Pearl bounds. Applying argmin_ϵ to Theorem 7.5.1 will provide a ϵ that will not narrow P(benefit) bounds beyond the Tian-Pearl bounds of (7.41). These ϵ-bounds will be an improvement on the Tian-Pearl bounds in (7.41) when $\epsilon < P(x,y') + P(x',y)$.

7.6 Examples

The following three examples demonstrate the value of refuting and confirming monotonicity in a mental health context.

7.6.1 Harmful Effects

A man is suing a pharmaceutical company claiming that he remained depressed *because* of their anti-depressant drug. While the company claims that the drug cannot worsen depression, the plaintiff asserts that he would've been cured on his own but the drug prolonged his depression. He presents the following data from an independent third party's observational

study. Does he have a case?

$$P(x) = 0.6, \tag{7.43}$$

$$P(y|x) = 0.3, \tag{7.44}$$

$$P(y|x') = 0.5. \tag{7.45}$$

This data shows that 30% of people choosing to take the drug recovered, while 50% of people choosing not to take the drug recovered. The pharmaceutical company conducts a Randomized Controlled Trial (RCT), suggesting that the reason people choosing their drug fared worse was because their willingness to incur the large expense of the drug was due to more severe depression where psychotherapy was ineffective. The RCT results showed the drug to be 11% effective in all categories measured, demonstrating that the above observational results were, in fact, biased due to confounding:

$$P(y_x) = 0.58,$$

$$P(y_{x'}) = 0.47.$$

The pharmaceutical company's experts further state that none of the drug's chemical mechanisms would make it possible for depression to be extended due to the drug itself. Our analysis, however, gives a different story.

Since $P(y) = P(y|x) \cdot P(x) + P(y|x') \cdot P(x') = 0.38$, the data fails the Monotonicity Necessity Test of Theorem 7.2.2. Specifically, $0.38 = P(y) \not\geqslant P(y_{x'}) = 0.47$. After plugging the data into Equation (3.23), the lower bound is $P(y_{x'}) - P(y) = 0.47 - 0.38 = 0.09$. Therefore, $P(\text{harm}) \geqslant 9\%$, contrary to the company's claim.

The interactive plot confirms this. After check-marking "Necessary" and "Observational data" and adjusting the probability sliders to match probabilities (7.43), (7.44), and (7.45), the coordinates $(P(y_x), P(y_{x'})) = (0.58, 0.47)$ point to the upper part of the purple band. Therefore, the pharmaceutical company is wrong and there is a risk of people staying depressed due to their drug.

The man has a strong claim in his lawsuit. Furthermore, market research may determine that, among potential customers, half of them would not purchase if they knew some people would remain depressed *because of* the drug. The anxiety induced by this knowledge could also make the drug less effective.

7.6.2 Confirming No Harm Claim

An ethical pharmaceutical company wants to proclaim that their drug to combat depression does not harm users. They believe that none of their users would simultaneously be cured without the drug *and* remain with depression after using the drug. However, they want to be responsible and confirm this before announcing anything. An RCT and observational study is conducted for this purpose, yielding the following:

$$\begin{aligned} P(x) &= 0.55, \\ P(y|x) &= 0.4, \\ P(y|x') &= 0.6, \\ P(y_x) &= 0.67, \\ P(y_{x'}) &= 0.27. \end{aligned}$$

One of the conditions of the sufficient test, Equation (7.11), is true:

$$\begin{aligned} P(x,y) + P(x',y') &= P(y|x) \cdot P(x) + P(y'|x') \cdot P(x') \\ &= P(y|x) \cdot P(x) + [1 - P(y|x')] \cdot [1 - P(x)] \\ &= 0.4 \\ &= \text{ATE} = P(y_x) - P(y_{x'}). \end{aligned}$$

Therefore, the pharmaceutical company can assure their customers of monotonicity. The interactive plot confirms this sufficient condition with the coordinates $(P(y_x), P(y_{x'})) = (0.67, 0.27)$ pointing to the bottom right of the purple band.

7.6.3 Improved Probability of Benefit

Remaining on our depression-drug theme, a third pharmaceutical company wishes to market their anti-depression drug as having a minimum 50% efficacy level for curing depression. They define efficacy as the proportion of people benefiting. The RCT they conducted for FDA approval yielded the following results:

$$P(y_x) = 0.55,$$
$$P(y_{x'}) = 0.46.$$

With only a paltry difference between experimental probabilities, $P(y_x)$ and $P(y_{x'})$, the ATE is $0.55 - 0.46 = 0.09$, which naively suggests low efficacy. Far below the hoped-for 50%. Even though this *average* treatment effect is low, the proportion of *individuals* benefiting may still be high. This is apparent when combing the RCT results above with the following observational study results:

$$\begin{aligned} P(x) &= 0.35, \\ P(y|x) &= 0.95, \\ P(y|x') &= 0.7, \\ P(y) &= 0.95 \cdot 0.35 + 0.7 \cdot 0.65 = 0.7875. \end{aligned}$$

It appears as though individuals are good at assessing whether they should consume this drug. Both the group choosing the drug and the group avoiding the drug fared better than both the treatment and control arms of the RCT. We can now compute bounds on $P(\text{benefit})$ using Equation (7.41), which represents the proportion of individuals benefiting from this drug:

$$\max\{0, 0.09, -0.2375, 0.3275\} \leqslant P(\text{benefit}) \leqslant \min\{0.55, 0.54, 0.5275, 0.5625\},$$
$$0.3275 \leqslant P(\text{benefit}) \leqslant 0.5275.$$

These results do not allow the pharmaceutical company to lay a legitimate claim to minimum 50% efficacy. However, they can claim *potentially up to* 52.75% efficacy. This is a vague claim, but may sway some hopeful depression sufferers to buy the drug.

Psychiatrists report that they believe many of their depressed patients are not getting better *because of* the drug, despite their belief that the drug is effective for an abundance of their other patients. The pharmaceutical company investigates and determines that the molecular mechanism does allow for some patients to be harmed by the drug. However, this mechanism only allows for a maximum of 24% of depressed people to be harmed. While this 0.24-limited harm is not ideal, the ATE is still positive, so psychiatrists and patients are largely amenable to continuing with the medication.

Unfortunately for the pharmaceutical company, 0.24-limited harm affects the $P(\text{benefit})$ bounds. Using Equation (7.42):

$$0.3275 \leqslant P(\text{benefit})_\epsilon \leqslant \min\{0.5275, 0.09 + 0.24\},$$
$$0.3275 \leqslant P(\text{benefit})_\epsilon \leqslant 0.33,$$

where $P(\text{benefit})_\epsilon$ is the probability of benefit incorporating ϵ-limited harm. $P(\text{benefit})$ is now shrunken to nearly a point estimate. The pharmaceutical company can no longer claim even a possibility of 50% efficacy.

7.7 Summary

Many reasoning tasks, such as unit selection, A/B testing, quasi-experimental econometrics, and, more generally, identification of Probabilities of Causation, benefit substantially from an assumption of monotonicity. This chapter showed how monotonicity can be detected (or refuted) from observational, experimental, or combined data. I then identify when monotonicity is definitely violated, when it definitely holds, and when it is undetermined. The consequences of monotonicity violations are further shown when the degree of violation is

limited.

CHAPTER 8

Selection Bias

8.1 Introduction

Selection bias occurs when a sample is not representative of the target population. It is induced by preferential selection of units for data analysis [BTP14]. There are two challenges selection bias poses to causal and statistical inference. First, selection bias can be difficult to avoid, or even detect. Second, analysis may be robust up to a certain level of selection bias. It can be important to ascertain how sensitive a result is to selection bias along with the severity of selection bias involved.

In order to tackle these challenges, we must first quantify the severity of selection bias. Then we obtain bounds on the selection bias involved in a sample. Finally, analysis can incorporate these bounds in its result.

(a) Z blocks path from S to Y (b) Confounder Z influences selection

Figure 8.1: Selection bias induced by Z

The selection mechanism can be represented in a causal graph with an S node. The DAG in figure 8.1a portrays a scenario where Z influences the outcome Y as well as whether the unit is selected into the sample. In order to block the selection mechanism's influence on the

outcome, the path from S to Y must be blocked. This can be accomplished by conditioning on or adjusting for Z. However, Z may be unobserved or even unknown.

8.2 Selection Bias Severity

The ATE is often the quantity of interest in an RCT. However, in an RCT, the sample's units would be conditioned on $S = 1$ (denoted by s for simplicity): $P(y_x|s) - P(y_{x'}|s)$. Let's call this CATE because it's conditional on having been selected. If there's a path from S to the outcome Y, as in figures 8.1a and 8.1b, then the difference between the population ATE and the CATE may be non-zero. This difference is how we can quantify SBS (selection bias severity):

$$\begin{aligned}\text{SBS} &\triangleq \text{ATE} - \text{CATE} \\ &= [P(y_x) - P(y_{x'})] - [P(y_x|s) - P(y_{x'}|s)].\end{aligned}$$

This number will measure the degree to which the ideal situation is violated. Bounds on SBS and bounds on the magnitude of SBS can be computed non-parametrically with five inputs: $P(y_x|s)$, $P(y_{x'}|s)$, $P(x)$, $P(y|x)$, and $P(y|x')$. The population-level counterfactual probabilities $P(y_x)$ and $P(y_{x'})$ are often estimated with RCTs as long as no selection bias occurs. An RCT yields estimates for $P(y_x|s)$ and $P(y_{x'}|s)$, so those quantities can be experimentally provided or computed using tools of causal inference. Population-level observational probabilities, $P(x)$, $P(y|x)$, and $P(y|x')$, are often much easier to estimate as no experiments are necessary.

If the ATE is identifiable from observed data, as in figure 8.1a, then SBS is directly calculable. However, the practical value of SBS, in this case, is limited as the RCT was unnecessary. On the other hand, figure 8.1b represents a scenario where the ATE is not identifiable when Z is unknown or unobserved. In this case, we can calculate bounds on the ATE.

Tian and Pearl [TP00] derived the following bounds on the causal effects $P(y_x)$ and $P(y_{x'})$:

$$P(x, y) \leqslant P(y_x) \leqslant 1 - P(x, y'),$$
$$P(x', y) \leqslant P(y_{x'}) \leqslant 1 - P(x', y').$$

These inequalities allow us to obtain the lower bound of ATE:

$$\begin{aligned} \text{ATE} &= P(y_x) - P(y_{x'}) \\ &\geqslant P(x, y) - [1 - P(x', y')] \\ &= P(x, y) + P(x', y') - 1. \end{aligned}$$

The upper bound of ATE is similarly derived:

$$\begin{aligned} \text{ATE} &\leqslant [1 - P(x, y')] - P(x', y) \\ &= 1 - P(x, y') - P(x', y) \\ &= P(x, y) + P(x', y'). \end{aligned}$$

The range of ATE, based solely on the probability distribution $P(X, Y)$, is always 1. The SBS is computed from ATE − CATE:

$$\text{SBS}_{\text{lower-bound}} = P(x, y) + P(x', y') - 1 - [P(y_x|s) - P(y_{x'}|s)],$$
$$\text{SBS}_{\text{upper-bound}} = P(x, y) + P(x', y') - [P(y_x|s) - P(y_{x'}|s)].$$

The quantity of interest is often the magnitude of selection bias. The absolute value of SBS is:

$$|\text{SBS}| \in \begin{cases} [\text{SBS}_{\text{lb}}, \text{SBS}_{\text{ub}}], & \text{if } \text{SBS}_{\text{lb}} > 0, \\ [-\text{SBS}_{\text{ub}}, -\text{SBS}_{\text{lb}}], & \text{if } \text{SBS}_{\text{ub}} < 0, \\ [0, \max(-\text{SBS}_{\text{lb}}, \text{SBS}_{\text{ub}})], & \text{if } 0 \in [\text{SBS}_{\text{lb}}, \text{SBS}_{\text{ub}}]. \end{cases} \tag{8.1}$$

These bounds are visualized at `https://sel-bias.vercel.app`.

Care must be taken when conditioning SBS on covariates. The causal effects of the ATE, $P(y_x)$ and $P(y_{x'})$, are not always equivalent to the interventional probabilities, $P(y|do(x))$ and $P(y|do(x'))$, when conditioning on a covariate Z [PGJ16, §4.3.1]. In particular, Z is a preintervention variable in the counterfactual probabilities $P(y_x|z)$ and $P(y_{x'}|z)$, while Z is a postintervention variable in the interventional probabilities, $P(y|do(x), z)$ and $P(y|do(x'), z)$. There's a difference if there exists a directed path from X to Z. Notably, the CATE, defined above, conditions on S. However, for most RCTs, the above counterfactual and interventional probabilities are equivalent as treatment assignment shouldn't influence inclusion in the sample. If it does, use the interventional probabilities.

8.3 Under Monotonicity

Tian and Pearl [TP00] derived narrower bounds on causal effects $P(y_x)$ and $P(y_{x'})$ under the condition of monotonicity:

$$P(y) \leqslant P(y_x) \leqslant 1 - P(x, y'),$$
$$P(x', y) \leqslant P(y_{x'}) \leqslant P(y).$$

This offers an opportunity for narrower bounds on SBS by narrowing the bounds on ATE:

$$\begin{aligned} \text{ATE} &= P(y_x) - P(y_{x'}) \\ &\geqslant P(y) - P(y) \\ &= 0. \end{aligned}$$

The upper bound of ATE remains the same as the non-parametric upper bound:

$$\begin{aligned} \text{ATE} &\leqslant [1 - P(x, y')] - P(x', y) \\ &= 1 - P(x, y') - P(x', y) \\ &= P(x, y) + P(x', y'). \end{aligned}$$

The range of ATE under monotonicity, based solely on the probability distribution $P(X,Y)$, is $P(x,y) + P(x',y')$. This range is always less than or equal to 1, thus providing superior SBS bounds to the non-parametric SBS bounds. The SBS bounds are:

$$\begin{aligned}
\text{SBS}_{\text{lower-bound}} &= -[P(y_x|s) - P(y_{x'}|s)] \\
&= P(y_{x'}|s) - P(y_x|s), \\
\text{SBS}_{\text{upper-bound}} &= P(x,y) + P(x',y') - [P(y_x|s) - P(y_{x'}|s)].
\end{aligned}$$

Because of monotonicity, $\text{SBS}_{\text{lower-bound}} \leqslant 0$. The magnitude of SBS is:

$$|\text{SBS}| \in \begin{cases} [\text{SBS}_{\text{lb}}, \text{SBS}_{\text{ub}}], & \text{if } P(y_{x'}|s) > P(y_x|s), \\ [-\text{SBS}_{\text{ub}}, -\text{SBS}_{\text{lb}}], & \text{if } P(y_x|s) > P(x,y) + P(x',y') + P(y_{x'}|s), \\ [0, \max(-\text{SBS}_{\text{lb}}, \text{SBS}_{\text{ub}})], & \text{otherwise.} \end{cases} \tag{8.2}$$

CHAPTER 9

Probabilities of Causation with Non-Binary Ordinal Outcomes

9.1 Introduction

This dissertation has presented methods to estimate PoCs more precisely than previously possible. These PoCs have a limitation, the outcomes are binary. This restriction either prevents these counterfactual probabilities from informing decisions effectively or it requires quantizing outputs to binary, which loses valuable information and can even cause us to answer the wrong question.

New methods have expanded on PoC bounds, particularly PN and P(benefit), to include non-binary outcomes [HFH19], though without incorporating observational data. Li and Pearl advanced PoC bounds to exploit observational data *and* work with both non-binary treatment and outcomes [LP24a]. In addition, Li and Pearl have applied these new bounds to the unit selection problem [LP24b].

This final chapter presents novel methods and an algorithm for more general bounds and identification of PoCs that encompass binary and non-binary ordinal outcomes, along with the consequences and redefining of the ATE, monotonicity, and unit selection. The restriction to ordinal outcomes allows for strong intuition behind the PoCs, monotonicity, and utility values, along with tractable bounds and identification strategies.

9.2 Probability of Benefit

We can define the Probability of Benefit with ordinal outcomes in an analogous way to Definition 3.5.1 and Pearl's Probability of Necessity and Sufficiency (PNS) [Pea09] with binary outcomes.

Definition 9.2.1 (Probability of Benefit)**.** *$Y \in (y_1, y_2, \ldots, y_n)$, where y_is are in increasing order. $X \in (x_1, x_2)$, where x_is are in increasing order.*

$$P(\textit{benefit}) \triangleq P(Y_{x_2} > Y_{x_1}) \tag{9.1}$$

$$= P\left(\bigvee_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i>j}} \left(y_{ix_2}, y_{jx_1}\right)\right) \tag{9.2}$$

$$= \sum_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i>j}} P\left(y_{ix_2}, y_{jx_1}\right). \tag{9.3}$$

Equation (9.1) has an intuitive appeal that, as we will see below, also lends itself well to extensions of the previous analyses. It is the probability that, for an individual, the outcome under treatment x_2 is better than the outcome under treatment x_1. The binary definition of the probability of benefiting from treatment, first described by Tian and Pearl [TP00], would be better served with this notation.

The disjunction in Equation (9.2) simply enumerates all counterfactual possibilities that comprise $P(Y_{x_2} > Y_{x_1})$. The final summation in Equation (9.3) follows from the fact that the counterfactual events in Equation (9.2) are mutually exclusive.

9.2.1 Bounds

We cannot simply replace $P(y_x, y'_{x'})$ with $P(y_{i_{x_2}}, y_{j_{x_1}})$ in the Tian-Pearl bounds for the binary-outcome P(benefit), as in

$$\max\left\{\begin{array}{c} 0, \\ P(y_{i_{x_2}}) + P(y_{j_{x_1}}) - 1, \\ P(y_i) - P(y_{i_{x_1}}), \\ P(y_{i_{x_2}}) - P(y_i), \end{array}\right\} \leqslant P(y_{i_{x_2}}, y_{j_{x_1}}) \leqslant \min\left\{\begin{array}{c} P(y_{i_{x_2}}), \\ P(y_{j_{x_1}}), \\ P(x_2, y_i) + P(x_1, y_j), \\ P(y_{i_{x_2}}) - P(y_{i_{x_1}}) + \\ P(x_2, y_j) + P(x_1, y_i) \end{array}\right\} \tag{9.4}$$

because Tian-Pearl bounds depend on y and y' being exhaustive of all outcome values.

Although the individual joint probability components of P(benefit) cannot be bounded according to (9.4), their sum can be bounded according to Theorem 9.2.2.

Theorem 9.2.2 (Bounds on P(benefit) with Ordinal Outcomes)**.**

$$P(\textit{benefit}) \geqslant \max\left\{\begin{array}{c} \sum_{\substack{1\leqslant i,j\leqslant n \\ s.t.\ i>j}} \max\left\{\begin{array}{c} 0, \\ P(y_{i_{x_2}}) + P(y_{j_{x_1}}) - 1 \end{array}\right\}, \\ P(y_{1_{x_1}}) - P(y_1), \\ P(y_n) - P(y_{n_{x_1}}), \\ P(y_1) - P(y_{1_{x_2}}), \\ P(y_{n_{x_2}}) - P(y_n), \\ \max_{1\leqslant i<n} \sum_{j=1}^{i} \left[P(y_{j_{x_1}}) - P(y_{j_{x_2}})\right] \end{array}\right\}, \tag{9.5}$$

$$P(\textit{benefit}) \leqslant \min\left\{\begin{array}{c} \sum_{\substack{1\leqslant i,j\leqslant n \\ s.t.\ i>j}} \min\left\{P(y_{i_{x_2}}), P(y_{j_{x_1}})\right\}, \\ 1 - P(x_2, y_1) - P(x_1, y_n), \\ 1 - P(y_{1_{x_2}}) - P(y_{n_{x_1}}) + P(x_2, y_1) + P(x_1, y_n), \\ 1 - \max_{1\leqslant i\leqslant n}\left\{\sum_{j=1}^{i}\left[P(y_{j_{x_2}}) - P(y_{j_{x_1}})\right] + P\left(y_{i_{x_1}}\right)\right\} \end{array}\right\}. \tag{9.6}$$

Compared with the Tian-Pearl bounds, there are more arguments to the outer min and max functions, and therefore more opportunities to narrow the bounds of P(benefit) with

ordinal outcomes where $|Y| \geqslant 3$. This does not mean that discretizing your outcome to a larger number of outcomes yields more precise bounds. This is both because the bounds are net necessarily narrower and also because we are asking a different question when P(benefit) is over a different number of outcomes. We will return to this matter at the end of this chapter.

The proofs in Appendix Sections 11.3.1.1 and 11.3.1.2 provide both qualitative conceptual proofs and mathematical proofs for both the lower and upper bounds.

9.2.2 Ternary Ordinal Outcome Example

Let us consider a fictional university organization, called GoGrad, with a mission to promote graduate school to undergraduate students. They devise a strategy of campus events and advertisements. Students would walk by their events at which point they could take an advertisement. At the same time, advertisements were randomly targeted to some students and explicitly prevented from other students.

There are $|Y| = 3$ outcomes for undergraduate students that were analyzed:

- y_1 = pursue Bachelor's degree
- y_2 = pursue Master's degree
- y_3 = pursue Ph.D. degree

The treatments were x_2 = received advertisement and x_1 = no advertisement.

Here is the data they collected after one year:

$$\begin{aligned}
P(y_{1_{x_1}}) &= 0.6, & P(y_{1_{x_2}}) &= 0.4,\\
P(y_{2_{x_1}}) &= 0.3, & P(y_{2_{x_2}}) &= 0.45,\\
P(y_{3_{x_1}}) &= 0.1, & P(y_{3_{x_2}}) &= 0.15,\\
P(y_1|x_1) &= 0.7, & P(y_1|x_2) &= 0.3,\\
P(y_2|x_1) &= 0.3, & P(y_2|x_2) &= 0.5,\\
P(y_3|x_1) &= 0, & P(y_3|x_2) &= 0.2,\\
P(x_1) &= 0.6, & P(x_2) &= 0.4.
\end{aligned}$$

The following additional probabilities are calculated for plugging into Equations (9.5) and (9.6):

$$\begin{aligned}
P(x_1, y_3) &= P(y_3|x_1) \cdot P(x_1) = 0 \cdot 0.6 = 0,\\
P(x_2, y_1) &= P(y_1|x_2) \cdot P(x_2) = 0.3 \cdot 0.4 = 0.12,\\
P(y_1) &= P(x_2, y_1) + P(y_1|x_1) \cdot P(x_1) = 0.12 + 0.7 \cdot 0.6 = 0.54,\\
P(y_3) &= P(x_1, y_3) + P(y_3|x_2) \cdot P(x_2) = 0 + 0.2 \cdot 0.4 = 0.08.
\end{aligned}$$

GoGrad wants to understand what proportion of students *benefited* from their efforts. In other words, how many students would have pursued a Bachelor's or Master's degree without their advertising and would have pursued a higher degree with their advertising:

$$P(\text{benefit}) = P\left[(y_{2_{x_2}}, y_{1_{x_1}}) \vee (y_{3_{x_2}}, y_{1_{x_1}}) \vee (y_{3_{x_2}}, y_{2_{x_1}})\right]$$

Let us now determine the lower bound on P(benefit) using Equation (9.5):

$$
\begin{aligned}
P(\text{benefit}) &\geqslant \max\left\{\begin{array}{c}
\max\{0, P(y_{2_{x_2}}) + P(y_{1_{x_1}}) - 1\} \\
+ \max\{0, P(y_{3_{x_2}}) + P(y_{1_{x_1}}) - 1\} \\
+ \max\{0, P(y_{3_{x_2}}) + P(y_{2_{x_1}}) - 1\}, \\
P(y_{1_{x_1}}) - P(y_1), \\
P(y_3) - P(y_{3_{x_1}}), \\
P(y_1) - P(y_{1_{x_2}}), \\
P(y_{3_{x_2}}) - P(y_3), \\
\max\{\sum_{j=1}^{1}\left[P(y_{j_{x_1}}) - P(y_{j_{x_2}})\right], \\
\sum_{j=1}^{2}\left[P(y_{j_{x_1}}) - P(y_{j_{x_2}})\right]\}
\end{array}\right\} \\
&= \max\left\{\begin{array}{c}
\max\{0, 0.05\} + \max\{0, -0.25\} + \max\{0, -0.55\}, \\
0.6 - 0.54, \\
0.08 - 0.1, \\
0.54 - 0.4, \\
0.15 - 0.08, \\
\max\{0.6 - 0.4, (0.6 - 0.4) + (0.3 - 0.45)\}
\end{array}\right\} \\
&= \max\Big\{0.05, 0.06, -0.02, 0.14, 0.07, 0.2\Big\} \\
&= 0.2.
\end{aligned}
$$

The upper bound on P(benefit) using Equation (9.6) is similarly calculated:

$$
\begin{aligned}
P(\text{benefit}) &\leqslant \min \left\{ \begin{array}{c} \max\{P(y_{2_{x_2}}), P(y_{1_{x_1}})\} \\ + \max\{P(y_{3_{x_2}}), P(y_{1_{x_1}})\} \\ + \max\{P(y_{3_{x_2}}), P(y_{2_{x_1}})\}, \\ 1 - P(x_2, y_1) - P(x_1, y_3), \\ 1 - P(y_{1_{x_2}}) - P(y_{3_{x_1}}) + P(x_2, y_1) + P(x_1, y_3), \\ 1 - \max\{\sum_{j=1}^{1} \left[P(y_{j_{x_2}}) - P(y_{j_{x_1}})\right] + P\left(y_{1_{x_1}}\right), \\ \sum_{j=1}^{2} \left[P(y_{j_{x_2}}) - P(y_{j_{x_1}})\right] + P\left(y_{2_{x_1}}\right)\} \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} \max\{0.45, 0.6\} + \max\{0.15, 0.6\} + \max\{0.15, 0.3\}, \\ 1 - 0.12 - 0, \\ 1 - 0.4 - 0.1 + 0.12 + 0, \\ 1 - \max\{0.4 - 0.6 + 0.6, (0.4 - 0.6 + 0.3) + (0.45 - 0.3 + 0.3)\} \end{array} \right\} \\
&= \min\{1.5, 0.88, 0.62, 1 - \max\{0.4, 0.55\}\} \\
&= 0.45.
\end{aligned}
$$

Therefore, the probability that any particular student benefits from receiving a GoGrad advertisement is between 20% and 45%.

9.2.3 Collapsing to Tian-Pearl Bounds

Bounds on P(benefit) with ordinal outcomes should supersede the Tian-Pearl bounds that are based on binary outcomes. As a base case verification of non-binary P(benefit) with ordinal outcomes, we can compare its bounds in Equations (9.5) and (9.6) when $|Y| = 2$ to the Tian-Pearl lower and upper bounds. The updated lower bound in Equation (9.5),

setting $n = 2$, is:

$$
\begin{aligned}
P(\text{benefit}) &\geqslant \max \left\{ \begin{array}{c} \sum_{\substack{1 \leqslant i,j \leqslant 2 \\ \text{s.t. } i>j}} \max \left\{ \begin{array}{c} 0, \\ P(y_{i x_2}) + P(y_{j x_1}) - 1 \end{array} \right\}, \\ P(y_{1 x_1}) - P(y_1), \\ P(y_2) - P(y_{2 x_1}), \\ P(y_1) - P(y_{1 x_2}), \\ P(y_{2 x_2}) - P(y_2), \\ \max_{2 \leqslant i \leqslant 2} \sum_{j=i}^{2} \left[P(y_{j x_2}) - P(y_{j x_1}) \right] \end{array} \right\} \\
&= \max \left\{ \begin{array}{c} \max \left\{ \begin{array}{c} 0, \\ P(y_{2 x_2}) + P(y_{1 x_1}) - 1 \end{array} \right\}, \\ 1 - P(y_{2 x_1}) - \left[1 - P(y_2) \right], \\ P(y_2) - P(y_{2 x_1}), \\ 1 - P(y_2) - \left[1 - P(y_{2 x_2}) \right], \\ P(y_{2 x_2}) - P(y_2), \\ P(y_{2 x_2}) - P(y_{2 x_1}) \end{array} \right\} \\
&= \max \left\{ \begin{array}{c} 0, \\ P(y_{2 x_2}) - P(y_{2 x_1}) \\ P(y_2) - P(y_{2 x_1}), \\ P(y_2) - P(y_{2 x_1}), \\ P(y_{2 x_2}) - P(y_2), \\ P(y_{2 x_2}) - P(y_2), \\ P(y_{2 x_2}) - P(y_{2 x_1}) \end{array} \right\} \\
&= \max \left\{ \begin{array}{c} 0, \\ P(y_{2 x_2}) - P(y_{2 x_1}) \\ P(y_2) - P(y_{2 x_1}), \\ P(y_{2 x_2}) - P(y_2), \end{array} \right\}. \qquad (9.7)
\end{aligned}
$$

Replacing y_2 with y, y_1 with y', x_2 with x, and x_1 with x' yields equivalent lower bounds between Equation (9.7) and Tian-Pearl's lower bound.

Similarly, the updated upper bound in Equation (9.6), setting $n = 2$, is:

$$
\begin{aligned}
P(\text{benefit}) &\leqslant \min \left\{ \begin{array}{c} \min\left\{P(y_{2_{x_2}}), P(y_{1_{x_1}})\right\}, \\ 1 - P(x_2, y_1) - P(x_1, y_2), \\ 1 - P(y_{1_{x_2}}) - P(y_{2_{x_1}}) + P(x_2, y_1) + P(x_1, y_2), \\ 1 + \min \left\{ \begin{array}{c} \sum_{j=1}^{2} \left[P(y_{j_{x_2}}) - P(y_{j_{x_1}})\right] - P(y_{1_{x_2}}), \\ P(y_{2_{x_2}}) - P(y_{2_{x_1}}) - P(y_{2_{x_2}}) \end{array} \right\} \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} P(y_{2_{x_2}}), \\ P(y_{1_{x_1}}), \\ P(x_2, y_2) + P(x_1, y_1), \\ P(y_{2_{x_2}}) - P(y_{2_{x_1}}) + P(x_2, y_1) + P(x_1, y_2), \\ 1 + \sum_{j=1}^{2} \left[P(y_{j_{x_2}}) - P(y_{j_{x_1}})\right] - P(y_{1_{x_2}}), \\ 1 - P(y_{2_{x_1}}) \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} P(y_{2_{x_2}}), \\ P(y_{1_{x_1}}), \\ P(x_2, y_2) + P(x_1, y_1), \\ P(y_{2_{x_2}}) - P(y_{2_{x_1}}) + P(x_2, y_1) + P(x_1, y_2), \\ P(y_{2_{x_2}}) + P(y_{1_{x_2}}) - P(y_{1_{x_1}}) + P(y_{2_{x_2}}) - P(y_{2_{x_1}}), \\ P(y_{1_{x_1}}) \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} P(y_{2_{x_2}}), \\ P(y_{1_{x_1}}), \\ P(x_2, y_2) + P(x_1, y_1), \\ P(y_{2_{x_2}}) - P(y_{2_{x_1}}) + P(x_2, y_1) + P(x_1, y_2), \\ 1 - 1 + P(y_{2_{x_2}}), \end{array} \right\}. \qquad (9.8)
\end{aligned}
$$

The last argument of the min function in Equation (9.8) is equivalent to the first argument and can be eliminated. As in the lower bound, this upper bound is equivalent to Tian-Pearl's

upper bound.

9.3 Probability of Harm

Using the same notation as in Section 9.2, we define the Probability of Harm.

Definition 9.3.1 (Probability of Harm)**.** *$Y \in (y_1, y_2, \ldots, y_n)$, where y_is are in increasing order. $X \in (x_1, x_2)$, where x_is are in increasing order.*

$$\begin{aligned} P(harm) &\triangleq P(Y_{x_2} < Y_{x_1}) \\ &= P\left(\bigvee_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i<j}} \left(y_{i x_2}, y_{j x_1}\right) \right) \\ &= \sum_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i<j}} P\left(y_{i x_2}, y_{j x_1}\right). \end{aligned}$$

Simply swapping x_1 and x_2 allows for an easy derivation of bounds on P(harm) from P(benefit).

Theorem 9.3.2 (Bounds on $P(\text{harm})$ with Ordinal Outcomes)**.**

$$P(harm) \geqslant \max \left\{ \begin{array}{c} \sum_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i>j}} \max \left\{ \begin{array}{c} 0, \\ P(y_{i_{x_1}}) + P(y_{j_{x_2}}) - 1 \end{array} \right\}, \\ \sum_{i=2}^{n} \left[\left(P(y_i) - P(y_{i_{x_2}}) \right) \right], \\ P(y_n) - P(y_{n_{x_2}}), \\ \sum_{i=2}^{n} \left[\left(P(y_{i_{x_1}}) - P(y_i) \right) \right], \\ P(y_{n_{x_1}}) - P(y_n), \\ \max_{2 \leqslant i \leqslant n} \sum_{j=i}^{n} \left[P(y_{j_{x_1}}) - P(y_{j_{x_2}}) \right] \end{array} \right\}, \tag{9.9}$$

$$P(harm) \leqslant \min \left\{ \begin{array}{c} \sum_{\substack{1 \leqslant i,j \leqslant n \\ s.t.\ i>j}} \min \left\{ P(y_{i_{x_1}}), P(y_{j_{x_2}}) \right\}, \\ \sum_{i=2}^{n} P(x_1, y_i) + \sum_{j=1}^{n-1} P(x_2, y_j), \\ \sum_{i=2}^{n} \left[P(y_{i_{x_1}}) - P(x_1, y_i) \right] + \sum_{i=1}^{n-1} \left[P(y_{i_{x_2}}) - P(x_2, y_j) \right], \\ 1 + \min_{1 \leqslant i \leqslant n} \left\{ \sum_{j=i}^{n} \left[P(y_{j_{x_1}}) - P(y_{j_{x_2}}) \right] - P(y_{i_{x_1}}) \right\} \end{array} \right\}. \tag{9.10}$$

9.4 Probability of Immunity

In previous literature, the binary outcome counterfactual $(y_x, y_{x'})$ has been referred to as immune, always-taker, and always cured [HS95]. Similarly, $(y'_x, y'_{x'})$ has been referred to as doomed, never-taker, and never cured. With non-binary ordinal outcomes, these adjectives are insufficient. There are $|Y|$ levels a unit can have with and without treatment, so we would need $|Y|$ adjectives. Instead, we will label every unit that is unaffected by treatment as immune.

Definition 9.4.1 (Probability of Immunity)**.** $Y \in (y_1, y_2, \ldots, y_n)$*, where* y_i*s are in increasing*

order. $X \in (x_1, x_2)$*, where* $x_i s$ *are in increasing order.*

$$\begin{aligned} P(\textit{immunity}) &\triangleq P(Y_{x_2} = Y_{x_1}) \\ &= P\left(\bigvee_{i=1}^{n} \left(y_{i_{x_2}}, y_{i_{x_1}}\right)\right) \\ &= \sum_{i=1}^{n} P\left(y_{i_{x_2}}, y_{i_{x_1}}\right). \end{aligned}$$

Let β_{lb}, β_{ub}, η_{lb}, and η_{ub} refer to the lower and upper bounds of P(benefit) and P(harm) in Equations (9.5), (9.6), (9.9), and (9.10), respectively.

Theorem 9.4.2 (Bounds on P(immunity) with Ordinal Outcomes)**.**

$$\max\{0, 1 - \beta_{ub} - \eta_{ub}\} \leqslant P(\textit{immunity}) \leqslant 1 - \beta_{lb} - \eta_{lb}.$$

This is easily derivable by recognizing that

$$\begin{aligned} P(Y_{x_2} = Y_{x_1}) &= 1 - P(Y_{x_2} > Y_{x_1}) - P(Y_{x_2} < Y_{x_1}) \\ &= 1 - P(\text{benefit}) - P(\text{harm}). \end{aligned}$$

9.5 ATE

The non-binary ordinal outcome probabilities of causation naturally accommodate both ordinal categorical outcomes and numerical outcomes, including those in which distances between numerical outcome values carry no meaningful interpretations. In contrast, the ATE is designed for numerical outcomes where differences between values directly reflect meaningful magnitudes. Although ATE is valuable in summarizing the effect of a treatment, its interpretation often becomes meaningless or nonsensical when dealing with categorical ordered outcomes. For example, consider numerical assignments to military ranks (e.g., 1 = Private, 2 = Corporal, 3 = Sergeant, 4 = Lieutenant). While the ordering conveys

progression in rank, the numerical differences between these ranks do not represent consistent or meaningful increments in advancement. Even when the outcome variable is numeric with meaningful differences between values, the ATE no longer aligns with the intuitive concept of causation as represented by the difference between probabilities $P(\text{benefit})$ and $P(\text{harm})$. The following mathematical derivation demonstrates why the equality ATE = $P(\text{benefit}) - P(\text{harm})$ breaks down when $|Y| > 2$.

$$
\begin{aligned}
\text{ATE} &= E[Y_{x_2} - Y_{x_1}] \\
&= E[Y_{x_2}] - E[Y_{x_1}] \\
&= \sum_{y \in Y} y \cdot P(Y_{x_2} = y) - \sum_{y \in Y} y \cdot P(Y_{x_1} = y) \\
&= \sum_{y \in Y} y \cdot \left[\sum_{k \in Y} P(y_{x_2}, y_{k x_1}) \right] - \sum_{y \in Y} y \cdot \left[\sum_{k \in Y} P(y_{k x_2}, y_{x_1}) \right] \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} P(y_{1x_2}, y_{1x_1}) & P(y_{1x_2}, y_{2x_1}) & \cdots & P(y_{1x_2}, y_{nx_1}) \\ P(y_{2x_2}, y_{1x_1}) & P(y_{2x_2}, y_{2x_1}) & \cdots & P(y_{2x_2}, y_{nx_1}) \\ \vdots & \vdots & \ddots & \vdots \\ P(y_{nx_2}, y_{1x_1}) & P(y_{nx_2}, y_{2x_1}) & \cdots & P(y_{nx_2}, y_{nx_1}) \end{bmatrix} \mathbf{1}_n \\
&\quad - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} P(y_{1x_2}, y_{1x_1}) & P(y_{1x_2}, y_{2x_1}) & \cdots & P(y_{1x_2}, y_{nx_1}) \\ P(y_{2x_2}, y_{1x_1}) & P(y_{2x_2}, y_{2x_1}) & \cdots & P(y_{2x_2}, y_{nx_1}) \\ \vdots & \vdots & \ddots & \vdots \\ P(y_{nx_2}, y_{1x_1}) & P(y_{nx_2}, y_{2x_1}) & \cdots & P(y_{nx_2}, y_{nx_1}) \end{bmatrix}^{\mathsf{T}} \mathbf{1}_n
\end{aligned}
$$

where $n = |Y|$. All of the component counterfactual probabilities that make up $P(\text{benefit})$ and $P(\text{harm})$ in Definitions 9.2.1 and 9.3.1 have differences of Y values as coefficients. These differences should equal 1 from $P(\text{benefit})$ components and -1 for $P(\text{harm})$ components in order for ATE = $P(\text{benefit}) - P(\text{harm})$ to hold. However, this system of linear equations has no solution. This is demonstrated below when $|Y| = 3$, with components of $P(\text{benefit})$

in teal and components of $P(\text{harm})$ in red:

$$\begin{aligned}
\text{ATE} = y_1 \cdot [P(y_{1_{x_2}}, y_{1_{x_1}}) + P(y_{1_{x_2}}, y_{2_{x_1}}) + P(y_{1_{x_2}}, y_{n_{x_1}})] \\
+ y_2 \cdot [P(y_{2_{x_2}}, y_{1_{x_1}}) + P(y_{2_{x_2}}, y_{2_{x_1}}) + P(y_{2_{x_2}}, y_{n_{x_1}})] \\
+ y_3 \cdot [P(y_{3_{x_2}}, y_{1_{x_1}}) + P(y_{3_{x_2}}, y_{2_{x_1}}) + P(y_{3_{x_2}}, y_{3_{x_1}})] \\
- y_1 \cdot [P(y_{1_{x_2}}, y_{1_{x_1}}) + P(y_{2_{x_2}}, y_{1_{x_1}}) + P(y_{3_{x_2}}, y_{1_{x_1}})] \\
- y_2 \cdot [P(y_{1_{x_2}}, y_{2_{x_1}}) + P(y_{2_{x_2}}, y_{2_{x_1}}) + P(y_{3_{x_2}}, y_{2_{x_1}})] \\
- y_3 \cdot [P(y_{1_{x_2}}, y_{3_{x_1}}) + P(y_{2_{x_2}}, y_{3_{x_1}}) + P(y_{3_{x_2}}, y_{3_{x_1}})] \\
= (y_1 - y_2) \cdot [P(y_{1_{x_2}}, y_{2_{x_1}})] + (y_1 - y_3) \cdot [P(y_{1_{x_2}}, y_{3_{x_1}})] \\
+ (y_2 - y_1) \cdot [P(y_{2_{x_2}}, y_{1_{x_1}})] + (y_2 - y_3) \cdot [P(y_{2_{x_2}}, y_{3_{x_1}})] \\
+ (y_3 - y_1) \cdot [P(y_{3_{x_2}}, y_{1_{x_1}})] + (y_3 - y_2) \cdot [P(y_{3_{x_2}}, y_{2_{x_1}})]. \qquad (9.11)
\end{aligned}$$

In order for $\text{ATE} = P(\text{benefit}) - P(\text{harm})$, the following must hold:

$$\begin{aligned}
y_1 - y_2 &= -1, \qquad (9.12) \\
y_1 - y_3 &= -1, \qquad (9.13) \\
y_2 - y_3 &= -1, \qquad (9.14) \\
y_2 - y_1 &= 1, \\
y_3 - y_1 &= 1, \\
y_3 - y_2 &= 1.
\end{aligned}$$

Equations (9.12), (9.13), and (9.14) ensure that the red components of $P(\text{harm})$ all have coefficient -1. Similarly, the remaining equations ensure that the teal components of $P(\text{benefit})$ all have coefficient 1.

$$\begin{aligned}
\text{ATE} = (-1) \cdot [P(y_{1_{x_2}}, y_{2_{x_1}})] + (-1) \cdot [P(y_{1_{x_2}}, y_{3_{x_1}})] \\
(1) \cdot [P(y_{2_{x_2}}, y_{1_{x_1}})] + (-1) \cdot [P(y_{2_{x_2}}, y_{3_{x_1}})] \\
(1) \cdot [P(y_{3_{x_2}}, y_{1_{x_1}})] + \quad (1) \cdot [P(y_{3_{x_2}}, y_{2_{x_1}})].
\end{aligned}$$

However, this system of linear equations has no solution. Subtracting Equation (9.14) from Equation (9.13) yields $y_1 - y_2 = 0$. This is in conflict with Equation (9.12). Therefore ATE $\neq P(\text{benefit}) - P(\text{harm})$ for non-binary ordinal outcomes.

Theorem 9.5.1 (ATE Function of PoCs)**.**

$$ATE = \sum_{1 \leqslant i,j \leqslant n} (y_i - y_j) \cdot P(y_{i x_2}, y_{j x_1}). \tag{9.15}$$

Equation (9.11) can be generalized to show the new relationship of ATE and PoCs:

$$\begin{aligned}
\text{ATE} = {} & (y_1 - y_2) \cdot p_{12} + (y_1 - y_3) \cdot p_{13} + \ldots + (y_1 - y_n) \cdot p_{1n} \\
& + (y_2 - y_1) \cdot p_{21} + (y_2 - y_3) \cdot p_{23} + \ldots + (y_2 - y_n) \cdot p_{2n} \\
& + \quad \vdots \quad + \quad \vdots \quad + \ddots + \quad \vdots \\
& + (y_n - y_1) \cdot p_{n1} + (y_n - y_2) \cdot p_{n2} + \ldots + (y_n - y_{n-1}) \cdot p_{n(n-1)} \\
= {} & \sum_{\substack{1 \leqslant i,j \leqslant n \\ \text{s.t. } i \neq j}} (y_i - y_j) \cdot P(y_{i x_2}, y_{j x_1})
\end{aligned}$$

where $p_{ij} = P(y_{i x_2}, y_{j x_1})$. Note that the summation in Equation (9.15) should be over all i and j such that $i \neq j$. However, $(y_i - y_j) = 0$ when $i = j$, so the condition $i \neq j$ is removed for simplicity.

Ju and Geng [JG10] proposed the following alternative to ATE to overcome some of the above challenges corresponding to counterfactual probabilities.

Definition 9.5.2 (Distributional Causal Effect (DCE))**.**

$$\begin{aligned}
DCE_k &= P(Y_{x_2} \geqslant y_k) - P(Y_{x_1} \geqslant y_k) \\
&= \sum_{j \geqslant k} P(y_{j x_2}) - \sum_{j \geqslant k} P(y_{j x_1}) \\
&= \sum_{j \geqslant k} \left[P(y_{j x_2}) - P(y_{j x_1}) \right].
\end{aligned}$$

However, DCE is not great as a summary treatment effect for two reasons. First, it does not incorporate values of outcome Y in cases where there are meaningful distances between

numeric values of Y. Second, there could be a mix of negative, zero, and positive average causal effects for different ordinal levels. For example, assume $|Y| = 3$ and

$$P(y_{2_{x_1}}) + P(y_{3_{x_1}}) \geqslant P(y_{2_{x_2}}) + P(y_{3_{x_2}}),$$
$$P(y_{3_{x_1}}) \leqslant P(y_{3_{x_2}}).$$

Then the 3 values for DCE are:

$$\begin{aligned}
\text{DCE}_1 &= 0, \\
\text{DCE}_2 &= P(y_{2_{x_2}}) + P(y_{3_{x_2}}) - \left[P(y_{2_{x_1}}) + P(y_{3_{x_1}})\right] \\
&< 0, \\
\text{DCE}_3 &= P(y_{3_{x_2}}) - P(y_{3_{x_1}}) \\
&> 0.
\end{aligned}$$

9.6 Monotonicity

With non-binary ordinal outcomes, monotonicity is defined as,

$$P(\text{harm}) = P(Y_{x_2} < Y_{x_1}) = 0. \tag{9.16}$$

Under this monotonicity assumption,

$$\begin{aligned}
P(\text{immunity}) &= 1 - P(Y_{x_2} > Y_{x_1}) - P(Y_{x_2} < Y_{x_1}) \\
&= 1 - P(Y_{x_2} > Y_{x_1}) \\
&= 1 - P(\text{benefit})
\end{aligned}$$

With binary outcomes, assuming monotonicity allows identification of $P(\text{benefit})$, $P(\text{harm})$, and $P(\text{immunity})$. Unfortunately, with non-binary ordinal outcomes, monotonicity, as defined in (9.16), is no longer sufficient for identification of those PoCs.

The reason monotonicity is sufficient to identify binary outcome PoCs is demonstrated

in the following relationship between the ATE, which is identifiably, and P(benefit):

$$\begin{aligned}\text{ATE} &= P(\text{benefit}) - P(\text{harm}) \\ &= P(\text{benefit}).\end{aligned}$$

However, under non-binary ordinal outcomes, ATE is no longer *necessarily* equivalent to the difference between P(benefit) and P(harm), as shown in Section (9.5).

All is not lost. We just need a stronger notion of monotonicity.

9.6.1 Monotonic Incremental Treatment Effect

Zhang et al. [ZGL24] defined *monotonic incremental treatment effect* for ordinal outcomes, which places an additional constraint on monotonicity. This assumption is reasonable to make in many real-world scenarios.

Definition 9.6.1 (Monotonic Incremental Treatment Effect (MITE))**.** *The treatment X does not decrease the level of Y and increases Y by at most one level, that is, $0 \leqslant Y_{x_2} - Y_{x_1} \leqslant 1$.*

Under the MITE assumption, the space of all counterfactual probabilities with binary treatment and ordinal outcomes becomes simplified to:

$$\begin{bmatrix} P(y_{1_{x_2}}, y_{1_{x_1}}) & 0 & 0 & \cdots & 0 & 0 \\ P(y_{2_{x_2}}, y_{1_{x_1}}) & P(y_{2_{x_2}}, y_{2_{x_1}}) & 0 & \cdots & 0 & 0 \\ 0 & P(y_{3_{x_2}}, y_{2_{x_1}}) & P(y_{3_{x_2}}, y_{3_{x_1}}) & \cdots & 0 & 0 \\ 0 & 0 & P(y_{4_{x_2}}, y_{3_{x_1}}) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & P(y_{n-1_{x_2}}, y_{n-1_{x_1}}) & 0 \\ 0 & 0 & 0 & \cdots & P(y_{n_{x_2}}, y_{n-1_{x_1}}) & P(y_{n_{x_2}}, y_{n_{x_1}}) \end{bmatrix} \tag{9.17}$$

The only non-zero entries are the diagonal and subdiagonal.

Using the Law of Total Probability (LoTP), summing the whole first row of entries in matrix (9.17) in order to sum out the event $y_{1_{x_1}}$ from $P(y_{1_{x_2}}, y_{1_{x_1}})$. This is because all entries

aside from $P(y_{1_{x_2}}, y_{1_{x_1}})$ in the first row are 0.

$$\begin{aligned} P(y_{1_{x_2}}, y_{1_{x_1}}) &= \sum_{i=1}^{n} P(y_{1_{x_2}}, y_{n_{x_1}}) \\ &= P(y_{1_{x_2}}). \end{aligned}$$

We are unable to use this same technique for the second row of (9.17) because there are two counterfactual PoCs. However, we sum the first *column* instead of row. There are still two counterfactual PoCs, but we just identified the first one.

$$\begin{aligned} P(y_{2_{x_2}}, y_{1_{x_1}}) &= \sum_{i=1}^{n} P(y_{n_{x_2}}, y_{1_{x_1}}) - P(y_{1_{x_2}}, y_{1_{x_1}}) \\ &= P(y_{1_{x_1}}) - P(y_{1_{x_2}}). \end{aligned}$$

Now that we have the first entry of the second row, we can sum the second row to identify $P(y_{2_{x_2}}, y_{2_{x_1}})$:

$$\begin{aligned} P(y_{2_{x_2}}, y_{2_{x_1}}) &= \sum_{i=1}^{n} P(y_{2_{x_2}}, y_{n_{x_1}}) - P(y_{2_{x_2}}, y_{1_{x_1}}) \\ &= P(y_{2_{x_2}}) - P(y_{1_{x_1}}) + P(y_{1_{x_2}}). \end{aligned}$$

We started with the first row, which identified its first element. Then we summed the first column, yielding its second element. Then we summed the second row, providing its second element. The second column can now be summed in order to identify $P(y_{3_{x_2}}, y_{2_{x_1}})$:

$$\begin{aligned} P(y_{3_{x_2}}, y_{2_{x_1}}) &= \sum_{i=1}^{n} P(y_{n_{x_2}}, y_{2_{x_1}}) - P(y_{2_{x_2}}, y_{2_{x_1}}) \\ &= P(y_{2_{x_1}}) - P(y_{2_{x_2}}) + P(y_{1_{x_1}}) - P(y_{1_{x_2}}). \end{aligned}$$

Continuing in this zig-zag fashion, we can recursively identify every counterfactual probability in (9.17):

$$\begin{aligned} P(y_{i_{x_2}}, y_{j_{x_1}}) = (i-j) \cdot \big[P(y_{j_{x_1}}) - P(y_{i-1_{x_2}}, y_{j_{x_1}})\big] \\ + (1-i+j) \cdot \big[P(y_{j_{x_2}}) - P(y_{i_{x_2}}, y_{j-1_{x_1}})\big] \end{aligned}$$

with the initial condition

$$P(y_{1_{x_2}}, y_{1_{x_1}}) = P(y_{1_{x_2}}),$$

where $i \in \{j, j+1\}$. If $i < j$ or $i > j+1$, then $P(y_{i_{x_2}}, y_{j_{x_1}}) = 0$. The closed form solution of this linear recurrence relation is shown in Theorem 9.6.2.

Theorem 9.6.2 (Identification of Probabilities of Causation under MITE)**.**

$$P(y_{i_{x_2}}, y_{j_{x_1}}) = \begin{cases} \sum_{k=1}^{j} P(y_{k_{x_2}}) - \sum_{k=1}^{j-1} P(y_{k_{x_1}}) & \textit{if } i = j \\ \sum_{k=1}^{j} \left[P(y_{k_{x_1}}) - P(y_{k_{x_2}}) \right] & \textit{if } i = j+1 \\ 0 & \textit{otherwise} \end{cases}$$

With binary outcomes, MITE is equivalent to monotonicity. We know that under monotonicity the binary outcome $P(\text{benefit}) = \text{ATE} = P(y_{2_{x_2}}) - P(y_{2_{x_1}})$. Theorem 9.6.2 agrees:

$$\begin{aligned} P(\text{benefit}) &= P(y_{2_{x_2}}, y_{1_{x_1}}) \\ &= P(y_{2_{x_2}}) - P(y_{2_{x_1}}). \end{aligned}$$

We can also verify that binary outcome $P(\text{immunity})$ and $P(\text{doom})$ are satisfied with Theorem 9.6.2. Since $P(\text{harm}) = 0$ under monotonicity, $P(\text{immunity}) = P(y_{2_{x_1}})$ and $P(\text{doom}) = P(y_{1_{x_2}})$. This is because under treatment x_1, the better outcome, y_2, is enjoyed by units that are harmed by treatment x_2 or immune, and, by assumption, no units are harmed.

$$\begin{aligned} P(\text{immunity}) &= P(y_{2_{x_2}}, y_{2_{x_1}}) \\ &= P(y_{2_{x_1}}). \end{aligned}$$

Similarly, under treatment x_2, the lesser outcome, y_1, is suffered by units that are harmed by treatment x_2 or doomed, and, by assumption, no units are harmed.

$$\begin{aligned} P(\text{doom}) &= P(y_{1_{x_2}}, y_{1_{x_1}}) \\ &= P(y_{1_{x_2}}). \end{aligned}$$

More generally, MITE is equivalent to monotonicity for binary treatment and outcomes.

9.6.2 Tests

The Monotonicity Necessity Test and the Monotonicity Sufficiency Test for binary outcomes are useful because we do not always know when monotonicity holds or when it cannot hold. The same challenge exists for MITE. Either MITE is true in the underlying data generating process but it is not justifiable, in which case a sufficiency test can be helpful, or MITE is not true in the underlying data generating process but it cannot be ruled out, in which case a necessary test can be helpful.

Theorem 9.6.3 (MITE Sufficiency Test)**.** *Y is monotonic, with incremental treatment effect, relative to X if*

$$\begin{aligned}
\eta_{ub} = 0, \ \textit{and} \\
[\forall i, j \in [n] \left(i > j+1 \implies P(y_{ix_2}) = 0 \vee P(y_{jx_1}) = 0\right), \ \textit{or} \\
P(x_2, y_1) + P(x_1, y_n) = 1, \ \textit{or} \\
1 + P(x_2, y_1) + P(x_1, y_n) = P(y_{1x_2}) + P(y_{nx_1}), \ \textit{or} \\
\max_{1 \leqslant i \leqslant n} \left\{ \sum_{j=1}^{i} \left[P(y_{jx_2}) - P(y_{jx_1})\right] + P\left(y_{ix_1}\right) \right\} = 1].
\end{aligned}$$

The first term ensures that $P(\text{harm}) = 0$. The second term ensures all other probabilities of causation where treatment increases Y by more than one level are 0. The expression comes from the Fréchet Inequality upper bound. The remaining terms come from $P(\text{benefit})$. While MITE does not require benefit to be 0, this does suffice and is added with an *or* condition.

Theorem 9.6.4 (MITE Necessity Test)**.** *Y is monotonic, with incremental treatment effect,*

relative to X only if

$$\begin{aligned}
\eta_{lb} = 0, \ \textit{and} \\
\forall i, j \in [n] \left(i > j + 1 \implies P(y_{i_{x_2}}) + P(y_{j_{x_1}}) = 1\right), \ \textit{and} \\
P(y_{1_{x_1}}) = P(y_1), \ \textit{and} \\
P(y_n) = P(y_{n_{x_1}}), \ \textit{and} \\
P(y_1) = P(y_{1_{x_2}}), \ \textit{and} \\
P(y_{n_{x_2}}) = P(y_n), \ \textit{and} \\
\max_{1 \leqslant i < n} \sum_{j=1}^{i} \left[P(y_{j_{x_1}}) - P(y_{j_{x_2}})\right] = 0.
\end{aligned}$$

The first term ensures it is possible that $P(\text{harm}) = 0$. The second term ensures all other probabilities of causation where treatment increases Y by more than one level are possibly 0. The expression comes from the Fréchet Inequality lower bound. For the same reason as the MITE Sufficiency Test, the remaining terms come from $P(\text{benefit})$.

9.7 Unit Selection

Li and Pearl [LP19] formulated the following benefit function, dependent on PoCs with binary outcomes.

$$f(\beta, \gamma, \theta, \delta) \triangleq \beta P(y_{2_{x_2}}, y_{1_{x_1}}) + \gamma P(y_{2_{x_2}}, y_{2_{x_1}}) + \theta P(y_{1_{x_2}}, y_{1_{x_1}}) + \delta P(y_{1_{x_2}}, y_{2_{x_1}}).$$

They showed that this benefit function is identifiable under monotonicity or gain equality.

Definition 9.7.1 (Gain Equality)**.**

$$\beta + \delta = \gamma + \theta. \tag{9.18}$$

9.7.1 Linear Benefit Function with Ordinal Outcomes

With non-binary ordinal outcomes, the Li-Pearl benefit function, $f(\beta, \gamma, \theta, \delta)$, needs to be generalized. With binary treatment and outcomes, there are 4 ways that individuals respond to the 2 treatments. They could be benefiters, harmed, immune, or doomed. The utility of selecting each of the counterfactual responder types needs to be expanded from 4 utility values corresponding to 4 responder types to $|Y|^{|X|} = |Y|^2$ utility values corresponding to $|Y|^2$ responder types.

A new benefit function is required.

Definition 9.7.2 (Linear Benefit Function with Ordinal Outcomes)**.**

$$b(\Upsilon) \triangleq \mathbf{1}_n^{\mathsf{T}} (\Upsilon \circ P) \mathbf{1}_n, \tag{9.19}$$

where $\Upsilon \in \mathbb{R}^{n \times n}$ is a matrix of utility values, $P \in \mathbb{R}^{n \times n}$ is a corresponding matrix of ordinal outcome probabilities of causation, and $\Upsilon \circ P$ is the element-wise Hadamard product of Υ and P.

Each element υ_{ij} of Υ corresponds to the utility of selecting a unit (y_{ix_2}, y_{jx_1}):

$$\Upsilon = \begin{bmatrix} \upsilon_{11} & \upsilon_{12} & \cdots & \upsilon_{1n} \\ \upsilon_{21} & \upsilon_{22} & \cdots & \upsilon_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{n1} & \upsilon_{n2} & \cdots & \upsilon_{nn} \end{bmatrix},$$

$$P = \begin{bmatrix} P(y_{1x_2}, y_{1x_1}) & P(y_{1x_2}, y_{2x_1}) & \cdots & P(y_{1x_2}, y_{nx_1}) \\ P(y_{2x_2}, y_{1x_1}) & P(y_{2x_2}, y_{2x_1}) & \cdots & P(y_{2x_2}, y_{nx_1}) \\ \vdots & \vdots & \ddots & \vdots \\ P(y_{nx_2}, y_{1x_1}) & P(y_{nx_2}, y_{2x_1}) & \cdots & P(y_{nx_2}, y_{nx_1}) \end{bmatrix}.$$

9.7.2 Utility Matrix Creation

Let us consider three different personas and how they would reason about the construction of a utility matrix Υ. First, Jack is a board member of GoGrad, the fictional university organization from Section 9.2.2. There are $|Y| = 3$ outcome possibilities, Bachelor's, Master's, and Ph.D. degrees. Therefore, $\Upsilon \in \mathbb{R}^{3^2 \times 3^2}$. The top row consists of numerical quantities that have some meaning to Jack as weights or preferences for counterfactual outcomes where students pursue Bachelor's degrees only after receiving a GoGrad advertisement. The first element is the value Jack places on advertising to a student who would have also pursued a Bachelor's degree only even if they had not received an advertisement. Jack figures GoGrad would have wasted an advertisement on them, so he choose a small negative dollar amount for υ_{11}. To the right is υ_{12}, where a student would have pursued a Master's degree without a GoGrad advertisement. Jack does not like this situation, the advertisement actually turned this student away from grad school. So Jack considers this a -\$20 loss to the organization. Next is υ_{13}, where a student would have pursued a Ph.D. without a GoGrad advertisement. This really infuriates Jack and he determines, since this is against everything GoGrad stands for, this is worth -\$60. The first row and the rest of the rows look like the following:

$$\Upsilon = \begin{bmatrix} -\$0.50 & -\$20 & -\$60 \\ \$15 & \$1 & -\$30 \\ \$40 & \$10 & -\$0.50 \end{bmatrix}$$

Notably, at \$40, Jack really values causing a student to pursue a Ph.D. when they would have otherwise pursued a Bachelor's degree. However, that benefit is not a harmful, in Jack's eyes, as causing the opposite (going from Ph.D. to Bachelor's). This asymmetric utility selection is entirely reasonable and is what separates great decision making from potentially severely sub-optimal decision making.

Second, consider Rebecca, head of the Ministry of Tourism for a beautiful mountain town. She has a great idea to incentivize travelers coming to her town with cryptocurrency in order to bring along friends and family. There are $|Y| = 3$ outcomes: traveler goes by

themselves, traveler brings 1 to 5 people, and traveler brings 6 or more people. If every traveler brought along 6 or more people, the town would be inundated and its reputation will be hurt. So Rebecca needs to be careful. She constructs her utility matrix as follows:

$$\Upsilon = \begin{bmatrix} -1 & -50 & -10 \\ 100 & -5 & 10 \\ -10 & -20 & -30 \end{bmatrix}$$

Rebecca's reasoning is not so obvious and there doesn't appear to be a clear pattern to her utilities. I will point out the notable points to help clarify her choices. The diagonal contains utilities for where the cryptocurrency incentive had no effect. They are negative because the incentive costs money and costs more money the more friends and family the traveler brings. This is why the utilities are negative and increase in magnitude towards the bottom right. The scenario in the second row and first column, where travelers would have came along without the incentive but bring 1 to 5 friends and family with the incentive, is worth the maximum value Rebecca awards. This is because it is the best outcome and exactly what her incentive program was designed for. The worst outcome is when a traveler was going to bring 1 to 5 people without the incentive but the incentive caused him to go alone. This is worth -50, which isn't as negative as 100 is positive, indicating that Rebecca cares more about a traveler benefiting her town than a traveler hurting her town. Finally, it may seem odd that Rebecca assigns a positive 10 to the situation when her incentive disencentivizes someone from bringing many people to now bring fewer people. Remember, she has to be cautious not to inundate her town with tourists.

Lastly, let us consider the STAR project (Chapter 2) utility matrix, Υ, with $|Y| = 4$ for the quaternary discretized outcomes. It is important that smaller class sizes benefit students' math scores, but even more important that they do not harm students' math scores. The Education Board of Directors decide to reward schools who take some intervention and benefit students. In the STAR case, the intervention was smaller class sizes. The following numbers are monetary rewards in thousands of dollars, written as single digits for space and

readability considerations.

$$\Upsilon = \begin{bmatrix} -2 & -3 & -5 & -10 \\ 2 & 1 & -1 & -5 \\ 4 & 3 & 1 & -1 \\ 5 & 4 & 2 & 0 \end{bmatrix}.$$

Notice the diagonal now increases in values when going towards the bottom right. This is because the Board feels that a poor math achieving student who remains a poor math achieving student with or without the intervention is not desirable. This desires a *penalty* of \$2,000. Whereas a below- or above-average math student who remains as such whether placed in a regular or small class is not a problem. The Board is completely fine with that and will even reward this with its smallest reward. This is because that student was not harmed and maybe was impossible to become a higher-achieving math student. On the other hand, a top performing math student who remains top performing with or without a small class is neutral. There is no reward or penalty. It was impossible to benefit this student with a smaller class as this student was already at the top. At least the student did not suffer with a small class. However, that is offset by the spot taken in the small class by a student who did not need it.

9.7.3 Identifiability

Under the assumption of MITE, all of P is identifiable, which makes $b(\Upsilon)$ identifiable. Since the new benefit function in Definition 9.7.2 requires $|Y|^2$ utility values, the gain equality in Equation (9.18) does not work to make the benefit function identifiable when $|Y| > 2$.

The utility values can still offer opportunities to identify $b(\Upsilon)$. Notice the columns and

the rows in the following Hadamard product of Υ and P:

$$\Upsilon \circ P = \begin{bmatrix} \upsilon_{11} \cdot P(y_{1x_2}, y_{1x_1}) & \upsilon_{12} \cdot P(y_{1x_2}, y_{2x_1}) & \cdots & \upsilon_{1n} \cdot P(y_{1x_2}, y_{nx_1}) \\ \upsilon_{21} \cdot P(y_{2x_2}, y_{1x_1}) & \upsilon_{22} \cdot P(y_{2x_2}, y_{2x_1}) & \cdots & \upsilon_{2n} \cdot P(y_{2x_2}, y_{nx_1}) \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{n1} \cdot P(y_{nx_2}, y_{1x_1}) & \upsilon_{n2} \cdot P(y_{nx_2}, y_{2x_1}) & \cdots & \upsilon_{n3} \cdot P(y_{nx_2}, y_{nx_1}) \end{bmatrix}.$$

If all of the utility values in a column are equal, that column collapses into an identifiable probability:

$$\upsilon_{1j} = \upsilon_{2j} = \ldots = \upsilon_{nj} \implies \sum_{i=1}^{n} \upsilon_{ij} \cdot P(y_{ix_2}, y_{j_{x_1}}) = \upsilon_{1j} \cdot P(y_{j_{x_1}}).$$

The same is true of the utility values in a row:

$$\upsilon_{i1} = \upsilon_{i2} = \ldots = \upsilon_{in} \implies \sum_{j=1}^{n} \upsilon_{ij} \cdot P(y_{ix_2}, y_{j_{x_1}}) = \upsilon_{i1} \cdot P(y_{ix_2}).$$

Unfortunately, it is unusual for the utility values to all be equal along a row or column. That would mean benefiting, harmed, and immune units all have the same utility, except at the edges of the matrix. Fortunately, the following is a simple test in $O(n^2)$ time to determine whether the utility values make $b(\Upsilon)$ identifiable, followed by a method to compute it.

Theorem 9.7.3 (Utility Equality Test (UET))**.**

$$\forall i, j \in \{2, 3, \ldots, n\} \left(\upsilon_{i1} - \upsilon_{ij} = \upsilon_{11} - \upsilon_{1j}\right). \tag{9.20}$$

To see why Theorem 9.7.3 works, we can split each utility value into the sum of two values: $\upsilon_{ij} = \alpha_i + \beta_j$, where α_is have the same value across each row and β_js have the same value across each column. For space considerations, let $p_{ij} = P(y_{ix_2}, y_{j_{x_1}})$:

$$\Upsilon \circ P = \begin{bmatrix} (\alpha_1 + \beta_1) \cdot p_{11} & (\alpha_1 + \beta_2) \cdot p_{12} & \cdots & (\alpha_1 + \beta_n) \cdot p_{1n} \\ (\alpha_2 + \beta_1) \cdot p_{21} & (\alpha_2 + \beta_2) \cdot p_{22} & \cdots & (\alpha_2 + \beta_n) \cdot p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ (\alpha_n + \beta_1) \cdot p_{n1} & (\alpha_n + \beta_2) \cdot p_{n2} & \cdots & (\alpha_n + \beta_n) \cdot p_{nn} \end{bmatrix}. \tag{9.21}$$

Notice that the sum of the diagonals of (9.21), multiplied by $n-1$ (the number of elements in a row or column minus 1 for the diagonal), equal the sum of the non-diagonals of (9.21):

$$\begin{aligned}(n-1)\cdot(\upsilon_{11}+\upsilon_{22}+\cdots+\upsilon_{nn}) &= \cancel{\upsilon_{11}}+\upsilon_{12}+\cdots+\upsilon_{1n}\\ &\quad+\upsilon_{21}+\cancel{\upsilon_{22}}+\cdots+\upsilon_{2n}\\ &\quad+\ \vdots\ +\ \vdots\ +\ddots+\ \vdots\\ &\quad+\upsilon_{n1}+\upsilon_{n2}+\cdots+\cancel{\upsilon_{nn}}\\ (n-1)\cdot(\alpha_1+\beta_1+\alpha_2+\beta_2+\cdots+\alpha_n+\beta_n) &= (n-1)\cdot\alpha_1+(n-1)\cdot\alpha_2\\ &\quad+(n-1)\cdot\beta_1+(n-1)\cdot\beta_2\\ &\quad+\quad\vdots\\ &\quad+(n-1)\cdot\alpha_n+(n-1)\cdot\beta_n\end{aligned}$$

This constitutes a necessary test, but it is not a sufficient test. The more strict constraint of Equation (9.20) is required.

The UET states that column differences are constant across rows and row differences are constant across columns. Consider an arbitrary difference between two υ elements on the same row:

$$\begin{aligned}\upsilon_{ij}-\upsilon_{ik} &= (\alpha_i+\beta_j)-(\alpha_i+\beta_k)\\ &= \beta_j-\beta_k.\end{aligned}$$

Since the difference in υ values on the same row do not depend on α_i, and therefore the specific row, that difference should be the same for every row between the same columns. The same invariant must hold for the difference between υs on the same column:

$$\begin{aligned}\upsilon_{ij}-\upsilon_{kj} &= (\alpha_i+\beta_j)-(\alpha_k+\beta_j)\\ &= \alpha_i-\alpha_k.\end{aligned}$$

We do not need to test both column differences across rows and row differences across columns, since one implies the other. Additionally, if the differences between the first column

and any column is constant across rows, that implies the difference between any pair of columns is constant across rows. Therefore, we only need to test the condition in (9.20). This is seen in the following derivation:

$$\begin{aligned} \upsilon_{ij} - \upsilon_{i1} &= \upsilon_{1j} - \upsilon_{11}, && \text{rows } i, 1\text{; same column differences} \\ \upsilon_{kj} - \upsilon_{k1} &= \upsilon_{1j} - \upsilon_{11}, && \text{rows } k, 1\text{; same column differences} \\ \upsilon_{ij} - \upsilon_{kj} &= \upsilon_{i1} - \upsilon_{k1}. && \text{columns } i, k\text{; same row differences} \end{aligned}$$

The first two equalities come from the UET condition and the third equality is the difference of the first two equalities. Note that there is never a need for the UET condition to test $i = 1$ or $j = 1$ because then the statement becomes a tautology.

When $|Y| = 2$, the UET simplifies to the Li-Pearl Gain Equality in Equation (9.18):

$$\begin{aligned} \upsilon_{21} - \upsilon_{22} &= \upsilon_{11} - \upsilon_{12}, \\ \beta &= \upsilon_{21}, \\ \gamma &= \upsilon_{22}, \\ \theta &= \upsilon_{11}, \\ \delta &= \upsilon_{12}, \\ \beta - \gamma &= \theta - \delta, \\ \gamma + \theta &= \beta + \delta. \end{aligned}$$

Note that the utility matrix Υ decomposition can easily be extended into non-binary treatment scenarios with an ($|X| = m$)-dimensional tensor comprised of m ($|Y| = n$)-dimensional vectors whose components are summed at each tensor location. The math in this chapter easily generalizes to accommodate this.

9.7.4 Identification Example

As an example, let us revisit the STAR project utility matrix constructed in Section 9.7.2. This does not pass the UET:

$$\begin{aligned} \upsilon_{13} - \upsilon_{14} &= -5 - -10 = 5, \\ \upsilon_{23} - \upsilon_{24} &= -1 - -5 = 4, \\ \upsilon_{13} - \upsilon_{14} &\neq \upsilon_{23} - \upsilon_{24}. \end{aligned}$$

Therefore, the STAR benefit is not identifiable. However, it is very close. Let us examine what happens if we simply change υ_{14} to -7 and υ_{24} to -3:

$$\Upsilon = \begin{bmatrix} -2 & -3 & -5 & -7 \\ 2 & 1 & -1 & -3 \\ 4 & 3 & 1 & -1 \\ 5 & 4 & 2 & 0 \end{bmatrix}. \tag{9.22}$$

The UET tells us that the benefit function $b(\Upsilon)$ is now identifiable since Υ passes condition (9.20):

$$\begin{aligned} 2 - 1 = -2 - -3 &\implies 1 = 1, \\ 2 - -1 = -2 - -5 &\implies 3 = 3, \\ 2 - -3 = -2 - -7 &\implies 5 = 5, \\ 4 - 3 = -2 - -3 &\implies 1 = 1, \\ 4 - 1 = -2 - -5 &\implies 3 = 3, \\ 4 - -1 = -2 - -7 &\implies 5 = 5, \\ 5 - 4 = -2 - -3 &\implies 1 = 1, \\ 5 - 2 = -2 - -5 &\implies 3 = 3, \\ 5 - 0 = -2 - -7 &\implies 5 = 5. \end{aligned}$$

It might seem strange why we should be able to simply change one's preference for certain outcomes or response types. In some circumstances this makes sense. In particular, the Tennessee Education Board wanted to incentivize schools based on the counterfactual responses of their students. However they cannot do this effectively without being able to point-estimate the associated PoCs. By simply lowering the penalty a little on two response types, the Board is able to achieve its objective.

9.7.5 Algorithm to Identify $b(\Upsilon)$ under Utility Equality

Given a utility matrix $\Upsilon \in \mathbb{R}^{n\times n}$ that passes the UET in Theorem 9.7.3, we want to find vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ such that $\upsilon_{ij} = a_i + b_j$. Then we can use $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to identify $\Upsilon \circ P$ and finally compute $b(\Upsilon)$. This process is performed in Algorithm 1.

The algorithm starts by simply assigning $\alpha_1 = 0$. Then all the β_js are easy to calculate, they are simply equal to $\upsilon_1 j$. We only need β_1 to find the α_is, α_i are equal to $\upsilon_{i1} - \beta_1$.

The reason we're allowed to assign $\alpha_1 = 0$, or any arbitrary value, is that you can always add any real value r to α_1 as long as you subtract r from β_1. In this way, $\alpha_1 + \beta_1 = \upsilon_{11}$ continues to hold. However, $\upsilon_{12}, \upsilon_{13}, \ldots, \upsilon_{1n}$ now need their corresponding β_is to be decremented by r as well. Similarly, because we decremented β_1 by r, $\upsilon_{21}, \upsilon_{31}, \ldots, \upsilon_{n1}$ need their corresponding α_is to be incremented by r. Ultimately, every α_i is incremented by r and every β_i is decremented by r, leaving all $\alpha_i + \beta_j = \upsilon_{ij}$ to remain true. Therefore, there are an infinite number of solutions for the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Note that although the UET is an $O(n^2)$ algorithm due to the need to check every element of an $n \times n$ matrix, the identification of the linear benefit function in Algorithm 1, once the UET is passed, is only a $O(n)$ algorithm. This can be computationally beneficial if we have a stable utility matrix, that passes the UET, applied to many different probabilistic scenarios. Step 2 in Algorithm 1 does call for checking the consistency of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and all elements of Υ, which is a $O(n^2)$ check, but this can be skipped when Υ passes the UET.

Algorithm 1 Identify $b(\Upsilon)$ under Utility Equality

Require:

- Utility matrix $\Upsilon \in \mathbb{R}^{n \times n}$.
- Marginal probability vectors $P_{x_2}, P_{x_1} \in \mathbb{R}^n$ of Probability of Causation matrix $P \in \mathbb{R}^{n \times n}$.

Ensure: Υ passes the UET in Theorem 9.7.3.

(1) Solve for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$, where $v_{ij} = \alpha_i + \beta_j$.
Set $\alpha_1 \leftarrow 0$
for $j = 1$ to n **do**
  $\beta_j \leftarrow v_{1j} - \alpha_1 = v_{1j}$
end for
for $i = 2$ to n **do**
  $\alpha_i \leftarrow v_{i1} - \beta_1$
end for
(2) Check consistency for all v_{ij} (optional).
for $i = 1$ to n **do**
  for $j = 1$ to n **do**
    if $v_{ij} \neq \alpha_i + \beta_j$ **then**
      Reject: no valid decomposition exists. **Stop**.
    end if
  end for
end for
(3) Compute the linear function of marginal probabilities of P.
Set $b \leftarrow \boldsymbol{\alpha}^\top P_{x_2} + \boldsymbol{\beta}^\top P_{x_1}$
return b
Accept: The decomposition exists with the found $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors and $b(\Upsilon)$ has been computed.

9.7.6 Simplification to Gain Equality

Algorithm 1, with $|Y| = 2$ and under the assumption of Gain Equality, simplifies to:

$$\begin{aligned}
\alpha_1 &= 0, \\
\beta_1 &= \upsilon_{11} = \theta, \\
\beta_2 &= \upsilon_{12} = \delta, \\
\alpha_2 &= \upsilon_{21} - \upsilon_{11} = \beta - \theta, \\
b(\Upsilon) &= (\beta - \theta) \cdot P(y_{2_{x_2}}) + \theta \cdot P(y_{1_{x_1}}) + \delta \cdot P(y_{2_{x_1}}) \\
&= (\beta - \theta) \cdot P(y_{2_{x_2}}) + \theta + (\delta - \theta) \cdot P(y_{2_{x_1}}). \qquad (9.23)
\end{aligned}$$

Equation (9.23) matches Li and Pearl's benefit function identification [Li and Pearl 2019, §5.1] when replacing $(\delta - \theta)$ by $(\gamma - \beta)$ due to Gain Equality.

9.7.7 Computation Example

Continuing with the quaternary discretized outcome STAR project utility matrix, Υ, in Equation (9.22), we can now compute a point-estimate of $b(\Upsilon)$ using Algorithm 1. The first step requires finding the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors:

$$\begin{aligned}
\alpha_1 &= 0, \\
\beta_1 &= \upsilon_{11} = -2, \\
\beta_2 &= \upsilon_{12} = -3, \\
\beta_3 &= \upsilon_{13} = -5, \\
\beta_4 &= \upsilon_{14} = -7, \\
\alpha_2 &= \upsilon_{21} - \upsilon_{11} = 2 - -2 = 4, \\
\alpha_3 &= \upsilon_{31} - \upsilon_{11} = 4 - -2 = 6, \\
\alpha_4 &= \upsilon_{41} - \upsilon_{11} = 5 - -2 = 7.
\end{aligned}$$

The next step in Algorithm 1 checks consistency for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and all elements of Υ. However, Υ has already passed the UET, so we can skip this step.

Finally, the benefit function $b(\Upsilon)$ can be computed as the sum of two matrix multiplications:

$$\begin{aligned}
b(\Upsilon) &= \boldsymbol{\alpha}^{\mathsf{T}} P_{x_2} + \boldsymbol{\beta}^{\mathsf{T}} P_{x_1} \\
&= \begin{bmatrix} 0 & 4 & 6 & 7 \end{bmatrix} \begin{bmatrix} P(y_{1x_2}) \\ P(y_{2x_2}) \\ P(y_{3x_2}) \\ P(y_{4x_2}) \end{bmatrix} + \begin{bmatrix} -2 & -3 & -5 & -7 \end{bmatrix} \begin{bmatrix} P(y_{1x_1}) \\ P(y_{2x_1}) \\ P(y_{3x_1}) \\ P(y_{4x_1}) \end{bmatrix} \\
&= 0(0.0062) + 4(0.2023) + 6(0.5290) + 7(0.2624) \\
&\quad + -2(0.0103) + -3(0.2437) + -5(0.5306) + -7(0.2154) \\
&= 0.9075.
\end{aligned}$$

The benefit is positive, so small classrooms are a good idea. The units and interpretation of $b(\Upsilon) = 0.9075$ depends on the units and interpretation of the utility matrix Υ.

Now let us see what the result would have been with binary discretized outcome. We will need a new utility matrix, which we will derive from the quaternary discretized utility matrix in Equation (9.22) by summing each of the four quadrants:

$$\Upsilon = \begin{bmatrix} -2 & -16 \\ 16 & 2 \end{bmatrix}. \tag{9.24}$$

We can immediately see this passes the UET because it satisfies Gain Equality in Equation (9.18):

$$16 + -16 = 2 + -2.$$

Let us find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\begin{aligned}
\alpha_1 &= 0, \\
\beta_1 &= \upsilon_{11} = -2, \\
\beta_2 &= \upsilon_{12} = -16, \\
\alpha_2 &= \upsilon_{21} - \upsilon_{11} = 16 - -2 = 18.
\end{aligned}$$

The overall benefit is then,

$$\begin{aligned}
b(\Upsilon) &= \begin{bmatrix} 0 & 18 \end{bmatrix} \begin{bmatrix} P(y_{1_{x_2}}) \\ P(y_{2_{x_2}}) \end{bmatrix} + \begin{bmatrix} -2 & -16 \end{bmatrix} \begin{bmatrix} P(y_{1_{x_1}}) \\ P(y_{2_{x_1}}) \end{bmatrix} \\
&= 0(0.2086) + 18(0.7914) + -2(0.2540) + -16(0.7460) \\
&= 1.8012.
\end{aligned}$$

It is interesting that the same STAR project, with the same data, yields two different benefits, 0.9075 when using a quaternary discretization and 1.8012 when using a binary discretization. This begs the question, why should we get almost double the benefit depending on how we discretize the outcome? Which result should we look at when making a decision?

We lose information when we discretize and when we compress the utility matrix. More importantly, we are answering a different benefit question when we consider different discretizations. This is similar to the additional information paradox of Section 4.3.4, where the explanation to the paradox is that we are answering different questions about bounds. For a binary ordinal outcome discretization, we are concerned with the benefit of overcoming some threshold. For a quaternary ordinal outcome discretization, we are concerned with how units would respond in 16 different ways, assuming binary treatment. The answer to the above question is to always use the largest possible discretization where the utility matrix can be decided and the Probability of Causation matrix can be estimated or causal effects can be estimated in the case that identification is possible.

9.7.8 Starting with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

We have seen how to assign utilities to counterfactual response types in Section 9.7.2, how to test if a utility assignment is identifiable with the UET in Theorem 9.7.3, and how to find the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors in order to compute the benefit function if utility equality is met in Algorithm 1. It is interesting to note that we do not need the UET or algorithm if we first start by creating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ instead of creating the utility matrix Υ.

There are two reasons to start with utility matrix creation versus $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ construction. First, we may not know that the utility values that define our decision strategy passes the UET and, therefore, allows the benefit function to be identified. If the UET fails, then it is impossible to construct corresponding $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors.

Second, it is easier to reason about the value of a unit that responds in a particular way to treatment 1 and in a particular way to treatment 2. To reason about $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, one must place values on causal effect probabilities under a particular treatment without regard to the response to the other treatment. For example, how would you place a utility value on treating someone who recovers when being administered this treatment? They might have recovered regardless of whether the treatment was administered and, for this utility, there is no information on how likely that is.

Some scenarios may lend themselves to effectively choosing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values. Consider what happens when $\boldsymbol{\alpha} = -\boldsymbol{\beta}$:

$$\begin{aligned} b(\Upsilon) &= \boldsymbol{\alpha}^{\mathsf{T}} P_{x_2} + \boldsymbol{\beta}^{\mathsf{T}} P_{x_1} \\ &= \boldsymbol{\alpha}^{\mathsf{T}} P_{x_2} - \boldsymbol{\alpha}^{\mathsf{T}} P_{x_1} \\ &= \boldsymbol{\alpha}^{\mathsf{T}} \left(P_{x_2} - P_{x_1} \right). \end{aligned}$$

The benefit function is then a version of ATE where the values of Y have been replaced with $\boldsymbol{\alpha}$. To go a step further, if the values of Y are numeric with meaningful distances, then $\boldsymbol{\alpha} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$ makes the benefit function exactly the ATE.

We now know how to create our utility matrix by starting with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and guaranteeing identifiability of the benefit function for a version of ATE or exactly the ATE. However, if someone wanted to set $\boldsymbol{\alpha} = -\boldsymbol{\beta}$, they could simply and directly use the ATE instead of the framework presented in this chapter. One valuable benefit they would receive is an examination of the constructed utility matrix. This would tell them the counterfactual response types they are really valuing, offering a deeper understanding of the often subjective values we place on things. Upon seeing the utility matrix, many people may wish to revise their utilities.

9.8 Continuous Outcome

The largest possible discretization of ordinal outcomes is when the cardinality of our outcome variable becomes infinite with either fine-grained values or by the nature of the outcome. In this case, we need a utility function, ω, of the number of treatment values analogous to the utility matrix Υ. Since this dissertation has focused on binary treatment, we can consider $\omega(\mu, \upsilon)$ to be a binary continuous function, where μ represents the outcome had treatment x_2 occurred and υ represents the outcome had treatment x_1 occurred. Similarly, $f(\mu, \upsilon)$ is defined as the counterfactual joint probability distribution function with the same meaning given to its two parameters.

Definition 9.8.1 (Continuous Benefit Function with Ordinal Outcomes)**.**

$$B(\omega) \triangleq \iint_{-\infty}^{\infty} \omega(\mu, \upsilon) \cdot f(\mu, \upsilon)\, d\mu\, d\upsilon.$$

It may be a challenge to specify the utility, preferences, or weights for $\omega(\mu, \upsilon)$ in a continuous mechanism, but many scenarios can allow this to be legitimately constructed. For example, an investment scenario can assign a financial utility as a function of the different outcomes under two different treatments.

The bigger challenge may be specifying the continuous joint probability distribution

function $f(\mu, \upsilon)$. Access to an underlying Structural Causal Model (SCM) can allow for this function to be constructed for point estimates at every μ and υ. Digital circuit analysis is an example scenario that can potentially identify $f(\mu, \upsilon)$.

CHAPTER 10

Conclusion

The research presented in this dissertation provides significant advancements in decision science by introducing and refining methods for personalized and situation-specific decision making using causal inference and counterfactual reasoning. Central to these advancements are several novel contributions.

Foundational counterfactual probabilities, called Probabilities of Causation (PoCs), were introduced and analyzed both mathematically and intuitively, offering techniques to estimate and apply them. Recognizing the historical difficulty of estimating these inherently counterfactual quantities, I demonstrated that, under reasonable assumptions and leveraging domain knowledge, causal structures, and covariate data, it is possible to derive sufficiently precise estimates to substantially improve decision quality.

I provided additional insight into refining decisions by demonstrating how the intention to act, as distinct from the action itself, can inform causal inference. Incorporating intentions and follow-through decisions as evidence offers deeper insights, helping further narrow counterfactual bounds and improve decisions.

Monotonicity, an assumption simplifying counterfactual estimation by constraining outcomes, was thoroughly examined. Necessary and sufficient conditions for monotonicity were tested, and the consequences of violations were analyzed. These results directly improved the accuracy and reliability of PoC estimates. Similarly, I addressed selection bias, providing methods to quantify, detect, and incorporate its effects into causal analyses, thereby enhancing the robustness of real-world decision making processes.

This dissertation culminated in generalized theorems that extended PoCs, unit selection, and decision making to scenarios with non-binary ordinal outcomes, moving beyond the binary success/failure paradigm. This cornerstone contribution derived new bounds, estimates, and conditions for point estimates, while adapting concepts like the Average Treatment Effect (ATE) and monotonicity to this more complex setting. A simple benefit function was introduced to guide optimal decision making, accompanied by an algorithm to compute overall utility.

The practical value of these methods was illustrated using a real-world dataset from Tennessee's STAR project. By combining observational and experimental data and applying these methodological innovations, clearer individual-level probabilities were produced, resulting in improved recommendations for educational decision making.

These contributions not only provide immediate tools for optimal decision making but also set the stage for future research and practical applications, including the transformative potential of integrating causal and counterfactual reasoning into AI systems.

10.1 Future Work

The findings and methods introduced in this dissertation pave the way for several promising future research directions:

- Explore richer combinations of covariates and mediators, beyond the scenarios investigated so far, to further narrow PoC bounds.
- Introduce new reasonable assumptions to achieve even narrower bounds on PoCs with both binary and non-binary ordinal outcomes.
- Formulate alternative monotonicity assumptions that allow for bound narrowing and point-estimating PoCs with non-binary ordinal outcomes.

- Narrow non-binary ordinal outcome PoC bounds utilizing the causal structure, using techniques in Chapter 4.
- Narrow non-binary ordinal outcome PoC bounds when the assumption of monotonicity is warranted, but MITE is not.
- Analyze consequences of partial violations of monotonicity and how that allows for bound narrowing of non-binary ordinal outcome PoCs.
- Extend the non-binary ordinal outcome framework to include non-binary ordinal treatments, in addition to outcomes. This would broaden the applicability of PoCs to more complex decision contexts.
- Narrow bounds on utility function based on the degree the UET fails
- Develop continuous benefit functions, building upon preliminary ideas discussed in Section 9.8, allowing for this decision making calculus to be applied to a wider range of scenarios.
- Create intuitive visualizations and interactive tools, similar to those available at `https://learn.ci/bounds.html`, to facilitate broader understanding and adoption of PoCs.
- Develop and publish software libraries to compute PoC bounds and integrate them with other Causes of Effects (CoE) algorithms and decision making algorithms. This would make these methods accessible to a wider audience of researchers and practitioners.

10.2 Artificial Intelligence

Achieving genuinely intelligent decision making in AI systems and agents demands the ability to reason about counterfactual scenarios and outcomes. Although state-of-the-art AI today is astonishing in prediction and pattern recognition, it often struggles to understand

causality or meaningfully incorporate counterfactual analysis. This is something us humans do instinctively, even as infants at a basic level [GG24].

To realize the full potential of AI, we should not wait until true causal and counterfactual reasoning become emergent properties. First, that day may never come. Second, even if current AI models can scale and be tweaked and trained to develop these capabilities on their own, we can dramatically accelerate progress now.

We need to integrate causal and counterfactual reasoning and decision making into the models. This might entail rearchitecting the fundamental framework and algorithms that comprise state of the art models. Instead of, or in addition to, that level of integration, embedding counterfactual thinking may take the form of careful selection and curation of pre-training data, fine-tuning models with a focus on causality, creating benchmarks that are only passable with a deep understanding of counterfactual reasoning, and ensuring the context of every interaction includes the math and algorithms of personalized and situation-specific decision making with counterfactual reasoning.

The methods presented and developed in this dissertation accomplish two goals. First they allow for optimal decision making with reasonable assumptions utilizing all available data. Second, they provide the math and algorithms for a machine to make the best possible choices.

REFERENCES

[ABB08] C. M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. "Tennessee's Student Teacher Achievement Ratio (STAR) project.", October 2008.

[BCI22] Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. "On Pearl's Hierarchy and the Foundations of Causal Inference (1st edition)." In Hector Geffner, Rita Dechter, and Joseph Halpern, editors, *Probabilistic and Causal Inference: the Works of Judea Pearl*, pp. 507–556. ACM Books, 2022.

[BP94] Alexander Balke and Judea Pearl. "Probabilistic evaluation of counterfactual queries." In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, pp. 230–237, Seattle, Washington, August 1994. AAAI Press.

[BP95] Alexander Balke and Judea Pearl. "Counterfactuals and policy analysis in structural models." In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, UAI'95, pp. 11–18, San Francisco, CA, USA, August 1995. Morgan Kaufmann Publishers Inc.

[BTP14] Elias Bareinboim, Jin Tian, and Judea Pearl. "Recovering from Selection Bias in Causal and Statistical Inference." *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**(1), June 2014. Number: 1.

[CFP24] Carlos Cinelli, Andrew Forney, and Judea Pearl. "A Crash Course in Good and Bad Controls." *Sociological Methods & Research*, **53**(3):1071–1104, August 2024. Publisher: SAGE Publications Inc.

[Che97] Patricia W. Cheng. "From covariation to causation: A causal power theory." *Psychological Review*, **104**(2):367–405, 1997. Place: US Publisher: American Psychological Association.

[DMM17] A Philip Dawid, Monica Musio, and Rossella Murtas. "The Probability of Causation." *Law, Probability and Risk*, **16**(4):163–179, December 2017.

[FM22] Andrew Forney and Scott Mueller. "Causal inference in AI education: A primer." *Journal of Causal Inference*, **10**(1):141–173, January 2022. Publisher: De Gruyter.

[GG24] Mariel K. Goddu and Alison Gopnik. "The development of human causal learning and reasoning." *Nature Reviews Psychology*, pp. 1–21, April 2024. Publisher: Nature Publishing Group.

[Gly13] Clark Glymour. "Psychological and Normative Theories of Causal Power and the Probabilities of Causes.", January 2013. arXiv:1301.7377 [cs].

[GP98] David Galles and Judea Pearl. "An Axiomatic Characterization of Causal Counterfactuals." *Foundations of Science*, **3**(1):151–182, January 1998.

[GP07] Sander Greenland and Judea Pearl. "Causal Diagrams." *Encyclopedia of Epidemiology*, pp. 149–156, 2007.

[GR86] Sander Greenland and James M Robins. "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology*, **15**(3):413–419, September 1986.

[Hal00] Joseph Y. Halpern. "Axiomatizing Causal Reasoning.", May 2000. arXiv:cs/0005030.

[HFH19] Emily J. Huang, Ethan X. Fang, Daniel F. Hanley, and Michael Rosenblum. "Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data." *Biometrics*, **75**(4):1228–1239, 2019.

[HR20] Miguel Hernán and James Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 1st edition edition, 2020.

[HS95] D. Heckerman and R. Shachter. "Decision-Theoretic Foundations for Causal Reasoning.", December 1995. arXiv:cs/9512104.

[IA94] Guido W. Imbens and Joshua D. Angrist. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, **62**(2):467–475, 1994. Publisher: [Wiley, Econometric Society].

[JG10] Chuan Ju and Zhi Geng. "Criteria for Surrogate end Points Based on Causal Distributions." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**(1):129–142, January 2010.

[KC11] Manabu Kuroki and Zhihong Cai. "Statistical Analysis of 'Probabilities of Causation' Using Co-variate Information." *Scandinavian Journal of Statistics*, **38**(3):564–577, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9469.2011.00730.x.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, July 2009.

[KFG89] M. J. Khoury, W. D. Flanders, S. Greenland, and M. J. Adams. "On the measurement of susceptibility in epidemiologic studies." *American Journal of Epidemiology*, **129**(1):183–190, January 1989.

[LDD18] Jiannan Lu, Peng Ding, and Tirthankar Dasgupta. "Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds." *Journal of Educational and Behavioral Statistics*, **43**(5):540–567, October 2018. Publisher: American Educational Research Association.

[LMP23] Ang Li, Scott Mueller, and Judea Pearl. "ϵ-Identifiability of Causal Quantities." January 2023.

[LP19] Ang Li and Judea Pearl. "Unit Selection Based on Counterfactual Logic." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1793–1799, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization.

[LP24a] Ang Li and Judea Pearl. "Probabilities of Causation with Nonbinary Treatment and Effect." *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**(18):20465–20472, March 2024. Number: 18.

[LP24b] Ang Li and Judea Pearl. "Unit Selection with Nonbinary Treatment and Effect." *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**(18):20473–20480, March 2024. Number: 18.

[LS18] Jeremy Labrecque and Sonja A. Swanson. "Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools." *Current Epidemiology Reports*, **5**(3):214–220, 2018.

[MLP22] Scott Mueller, Ang Li, and Judea Pearl. "Causes of Effects: Learning Individual Responses from Population Data." In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2712–2718, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization.

[MP19] Scott Mueller and Judea Pearl. "Fréchet Inequalities – Visualization, Applications, and History.", November 2019.

[MP20] Scott Mueller and Judea Pearl. "Which Patients are in Greater Need: A counterfactual analysis with reflections on COVID-19.", April 2020.

[MP21] Scott Mueller and Judea Pearl. "Personalized Decision Making.", April 2021.

[MP23a] Scott Mueller and Judea Pearl. "Monotonicity: Detection, Refutation, and Ramification." August 2023.

[MP23b] Scott Mueller and Judea Pearl. "Personalized Decision Making under Concurrent-Controlled RCT Data.", March 2023.

[MP23c] Scott Mueller and Judea Pearl. "Personalized decision making – A conceptual introduction." *Journal of Causal Inference*, **11**(1), January 2023. Publisher: De Gruyter.

[MP24] Scott Mueller and Judea Pearl. "The Meaning of 'Harm' in Personalized Medicine – An Alternative Perspective." *American Journal of Epidemiology*, November 2024.

[Mue21] Scott Allen Mueller. *Estimating Individualized Causes of Effects by Leveraging Population Data*. PhD thesis, UCLA, 2021.

[MW14] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, Cambridge, 2 edition, 2014.

[Pea93] Judea Pearl. "Aspects of Graphical Models Connected With Causality." *Proceedings of the 49th Session of the International Statistical Institute, Italy*, pp. 399–401, 1993.

[PEA95] JUDEA PEARL. "Causal diagrams for empirical research." *Biometrika*, **82**(4):669–688, December 1995.

[Pea99] Judea Pearl. "Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification." *Synthese*, **121**(1):93–149, November 1999.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009.

[Pea10] Judea Pearl. "On the consistency rule in causal inference: axiom, definition, assumption, or theorem?" *Epidemiology (Cambridge, Mass.)*, **21**(6):872–875, November 2010.

[Pea11] Judea Pearl. "Principal Stratification – a Goal or a Tool?" *The International Journal of Biostatistics*, **7**(1):1–13, March 2011. Publisher: De Gruyter.

[Pea13] Judea Pearl. "Understanding Simpson's Paradox.", September 2013.

[Pea15] Judea Pearl. "Causes of Effects and Effects of Causes." *Sociological Methods & Research*, **44**(1):149–164, February 2015. Publisher: SAGE Publications Inc.

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, March 2016.

[PP10] Judea Pearl and Azaria Paz. "Confounding Equivalence in Causal Inference." In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 433–441, Arlington, Virginia, USA, July 2010. AUAI Press.

[Res] Straits Research. "Global Nutritional Supplements Market Size to Hit USD 816.57 billion by 2033.".

[RR13] T. Richardson and J. Robins. "Single World Intervention Graphs : A Primer.", 2013.

[Rub74] Donald B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, **66**(5):688–701, 1974. Place: US Publisher: American Psychological Association.

[Sen10] Stephen Senn. "Control in Clinical Trials." *Proceedings of the Eighth International Conference on Teaching Statistics*, 2010.

[SGS01] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, January 2001.

[SP08] Ilya Shpitser and Judea Pearl. "Complete Identification Methods for the Causal Hierarchy." *Journal of Machine Learning Research*, **9**(64):1941–1979, 2008.

[SP12] Ilya Shpitser and Judea Pearl. "What Counterfactuals Can Be Tested.", June 2012. arXiv:1206.5294 [cs].

[TP00] Jin Tian and Judea Pearl. "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence*, **28**(1):287–313, October 2000.

[WJB90] Elizabeth Word, John Johnston, Helen Pate Bain, B. DeWayne Fulton, Jayne Boyd Zaharias, Charles M. Achilles, Martha Nannette Lintz, John Folger, and Carolyn Breda. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project." Technical report, Tennessee State Department of Education, 1990. ERIC Number: ED320692.

[ZAC95] Jayne B. Zaharias, C. M. Achilles, and Van A. Cain. "The Effect of Random Class Assignment on Elementary Students' Reading and Mathematics Achievement." *Research in the Schools*, **2**(2):7–14, 1995. ERIC Number: EJ571155.

[ZGL24] Chao Zhang, Zhi Geng, Wei Li, and Peng Ding. "Identifying and bounding the probability of necessity for causes of effects with ordinal outcomes.", November 2024. arXiv:2411.01234 [math].

[ZLM24] Chi Zhang, Ang Li, Scott Mueller, and Rumen Iliev. "Causal AI Framework for Unit Selection in Optimizing Electric Vehicle Procurement." Vancouver, Canada, February 2024.

CHAPTER 11

Appendices

11.1 Appendix for Chapter 3

11.1.1 Bounds of $P(\textbf{harm})$

The bounds of $P(\text{harm})$ in Equation (3.23) come from

$$\begin{aligned} P(\text{harm}) &= P(\text{benefit}) - \text{ATE} \\ &= P(\text{benefit}) - [P(y_x) - P(y_{x'})] \\ &= P(\text{benefit}) - P(y_x) + P(y_{x'}) \end{aligned}$$

and the bounds of $P(\text{benefit})$:

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \leqslant P(\text{benefit}) \leqslant \min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) \\ + \ P(x,y') + P(x',y) \end{array} \right\}.$$

We can now bound $P(\text{harm})$:

$$
\begin{aligned}
P(\text{harm}) &\geqslant \max \left\{ \begin{array}{c} 0 - P(y_x) + P(y_{x'}), \\ P(y_x) - P(y_{x'}) - P(y_x) + P(y_{x'}), \\ P(y) - P(y_{x'}) - P(y_x) + P(y_{x'}), \\ P(y_x) - P(y) - P(y_x) + P(y_{x'}) \end{array} \right\} \\
&= \max \left\{ \begin{array}{c} P(y_{x'}) - P(y_x), \\ 0, \\ P(y) - P(y_x), \\ P(y_{x'}) - P(y) \end{array} \right\}, \\
P(\text{harm}) &\leqslant \min \left\{ \begin{array}{c} P(y_x) - P(y_x) + P(y_{x'}), \\ P(y'_{x'}) - P(y_x) + P(y_{x'}), \\ P(x,y) + P(x',y') - P(y_x) + P(y_{x'}), \\ P(y_x) - P(y_{x'}) \\ + \ P(x,y') + P(x',y) - P(y_x) + P(y_{x'}) \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} P(y_{x'}), \\ [1 - P(y_{x'})] - P(y_x) + P(y_{x'}), \\ P(y_{x'}) - P(y_x) \\ + \ P(x,y) + P(x',y'), \\ P(x,y') + P(x',y) \end{array} \right\} \\
&= \min \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x,y') + P(x',y), \\ P(y_{x'}) - P(y_x) \\ + \ P(x,y) + P(x',y') \end{array} \right\}.
\end{aligned}
$$

11.1.2 Relationships of P(immunity) and P(doom) to Benefit and Harm

Equations (3.24) and (3.25) are derived as follows:

$$\begin{aligned} P(\text{immunity}) &= P(y_x, y_{x'}) \\ &= P(y_x, y_{x'}) + [P(y_x, y'_{x'}) - P(y_x, y'_{x'})] \\ &= P(y_x) - P(y_x, y'_{x'}) \qquad (11.1) \\ &= P(y_x) - P(\text{benefit}), \\ P(\text{doom}) &= P(y'_x, y'_{x'}) \\ &= P(y'_x, y'_{x'}) + [P(y'_x, y_{x'}) - P(y'_x, y_{x'})] \\ &= P(y'_x) - P(y'_x, y_{x'}) \qquad (11.2) \\ &= P(y'_x) - P(\text{harm}). \end{aligned}$$

The LoTP allows for Equations (11.1) and (11.2).

11.1.3 Bounds of $P(\textbf{immunity})$ and $P(\textbf{harm})$

The lower bound of Equation (3.26) is derived using Equation (3.24) as follows:

$$
\begin{aligned}
P(\text{immunity}) \geqslant P(y_x) - \min & \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) \\ +\ P(x,y') + P(x',y) \end{array} \right\} \\
= \max & \left\{ \begin{array}{c} P(y_x) - P(y_x), \\ P(y_x) - P(y'_{x'}), \\ P(y_x) - [P(x,y) + P(x',y')], \\ P(y_x) - [P(y_x) - P(y_{x'}) \\ \quad +\ P(x,y') + P(x',y)] \end{array} \right\} \\
= \max & \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y'_{x'}), \\ P(y_x) - P(x,y) - P(x',y') \\ P(y_{x'}) - P(x,y') - P(x',y) \end{array} \right\}.
\end{aligned}
$$

The upper bound of Equation (3.26) is derived using Equation (3.24) as follows:

$$
\begin{aligned}
P(\text{immunity}) \leqslant P(y_x) - \max & \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \\
= \min & \left\{ \begin{array}{c} P(y_x) - 0, \\ P(y_x) - [P(y_x) - P(y_{x'})], \\ P(y_x) - [P(y) - P(y_{x'})], \\ P(y_x) - [P(y_x) - P(y)] \end{array} \right\} \\
= \min & \left\{ \begin{array}{c} P(y_x), \\ P(y_{x'}), \\ P(y_x) + P(y_{x'}) - P(y), \\ P(y) \end{array} \right\}
\end{aligned}
$$

Similarly, the lower bound of (3.27) is derived using Equation (3.25) as follows:

$$
\begin{aligned}
P(\text{doom}) \geqslant P(y'_x) - \min & \left\{ \begin{array}{c} P(y_{x'}), \\ P(y'_x), \\ P(x,y') + P(x',y), \\ P(y_{x'}) - P(y_x) \\ \quad + P(x,y) + P(x',y') \end{array} \right\} \\
= \max & \left\{ \begin{array}{c} P(y'_x) - P(y_{x'}), \\ P(y'_x) - P(y'_x), \\ P(y'_x) - [P(x,y') + P(x',y)], \\ P(y'_x) - [P(y_{x'}) - P(y_x) \\ \quad + P(x,y) + P(x',y')] \end{array} \right\} \\
= \max & \left\{ \begin{array}{c} P(y'_x) - P(y_{x'}), \\ 0, \\ P(y'_x) - P(x,y') - P(x',y) \\ P(y'_x) + P(y_x) - P(y_{x'}) \\ \quad - P(x,y) - P(x',y') \end{array} \right\} \\
= \max & \left\{ \begin{array}{c} 0, \\ P(y'_x) - P(y_{x'}), \\ P(y'_x) - P(x,y') - P(x',y) \\ P(y'_{x'}) - P(x,y) - P(x',y') \end{array} \right\}.
\end{aligned}
$$

The last argument to max in the final equality is a result of $P(y'_x)+P(y_x) = 1$ and $1-P(y_{x'}) = P(y'_{x'})$.

Finally, the upper bound of (3.27) is derived using Equation (3.25) as follows:

$$P(\text{doom}) \leqslant P(y'_x) - \max \left\{ \begin{array}{c} 0, \\ P(y_{x'}) - P(y_x), \\ P(y) - P(y_x), \\ P(y_{x'}) - P(y) \end{array} \right\}$$

$$= \min \left\{ \begin{array}{c} P(y'_x) - 0, \\ P(y'_x) - [P(y_{x'}) - P(y_x)], \\ P(y'_x) - [P(y) - P(y_x)], \\ P(y'_x) - [P(y_{x'}) - P(y)] \end{array} \right\}$$

$$= \min \left\{ \begin{array}{c} P(y'_x), \\ P(y'_x) + P(y_x) - P(y_{x'}), \\ P(y'_x) + P(y_x) - P(y), \\ P(y'_x) - P(y) + P(y_{x'}) \end{array} \right\}$$

$$= \min \left\{ \begin{array}{c} P(y'_x), \\ P(y'_{x'}), \\ P(y'), \\ P(y'_x) + P(y'_{x'}) - P(y') \end{array} \right\}.$$

The second-to-last argument to max in the final equality is a result of $P(y'_x) + P(y_x) = 1$ and $1 - P(y) = P(y')$.

11.2 Appendix for Chapter 4

11.2.1 $P(\textbf{harm})$ Under Exogeneity

The ATE under exogeneity becomes:

$$\begin{aligned} \text{ATE} &= P(y_x) - P(y_{x'}) \\ &= P(y|x) - P(y|x'). \end{aligned} \tag{11.3}$$

We can now derive $P(\text{harm})$ under exogeneity in Equation (4.9) from $P(\text{benefit})$ under exogeneity through their relationship with the ATE in Equation (3.22) and Equation (11.3):

$$\max\{0, P(y|x) - P(y|x')\} \leqslant P(\text{benefit}) \leqslant \min\{P(y|x), P(y'|x')\},$$

$$\begin{smallmatrix}\max\{0,P(y|x)-P(y|x')\}\\ -[P(y_x)-P(y_{x'})]\end{smallmatrix} \leqslant \begin{smallmatrix}P(\text{benefit})\\ -[P(y_x)-P(y_{x'})]\end{smallmatrix} \leqslant \begin{smallmatrix}\min\{P(y|x),P(y'|x')\}\\ -[P(y_x)-P(y_{x'})]\end{smallmatrix},$$

$$\begin{smallmatrix}\max\{0,P(y|x)-P(y|x')\}\\ -P(y|x)+P(y|x')\end{smallmatrix} \leqslant P(\text{harm}) \leqslant \begin{smallmatrix}\min\{P(y|x),P(y'|x')\}\\ -P(y|x)+P(y|x')\end{smallmatrix},$$

$$\max\{P(y|x') - P(y|x), 0\} \leqslant P(\text{harm}) \leqslant \min\{P(y|x'), 1 - P(y|x)\},$$

$$\max\{0, P(y|x') - P(y|x)\} \leqslant P(\text{harm}) \leqslant \min\{P(y'|x), P(y|x')\}.$$

11.2.2 $P(\textbf{immunity})$ Under Exogeneity

We can derive $P(\text{immunity})$ Under Exogeneity in Equation (4.10) using Equation (3.24):

$$\max\{0, P(y|x) - P(y|x')\} \leqslant P(\text{benefit}) \leqslant \min\{P(y|x), P(y'|x')\},$$

$$-\min\{P(y|x), P(y'|x')\} \leqslant -P(\text{benefit}) \leqslant -\max\{0, P(y|x) - P(y|x')\},$$

$$P(y_x) - \min\{P(y|x), P(y'|x')\} \leqslant P(\text{immunity}) \leqslant P(y_x) - \max\{0, P(y|x) - P(y|x')\},$$

$$P(y|x) + \max\{-P(y|x), -P(y'|x')\} \leqslant P(\text{immunity}) \leqslant P(y|x) + \min\{0, P(y|x') - P(y|x)\},$$

$$\max\{0, P(y|x) - P(y'|x')\} \leqslant P(\text{immunity}) \leqslant \min\{P(y|x), P(y|x')\}.$$

11.2.3 $P(\textbf{doom})$ Under Exogeneity

We can derive $P(\text{doom})$ Under Exogeneity in Equation (4.11) using Equation (3.25):

$$\max\{0, P(y|x') - P(y|x)\} \leqslant P(\text{harm}) \leqslant \min\{P(y'|x), P(y|x')\},$$

$$-\min\{P(y'|x), P(y|x')\} \leqslant -P(\text{harm}) \leqslant -\max\{0, P(y|x') - P(y|x)\},$$

$$P(y'_x) - \min\{P(y'|x), P(y|x')\} \leqslant P(\text{doom}) \leqslant P(y'_x) - \max\{0, P(y|x') - P(y|x)\},$$

$$P(y'|x) + \max\{-P(y'|x), -P(y|x')\} \leqslant P(\text{doom}) \leqslant P(y'|x) + \min\{0, P(y|x) - P(y|x')\},$$

$$\max\{0, P(y'|x) - P(y|x')\} \leqslant P(\text{doom}) \leqslant \min\{P(y'|x), P(y'|x')\}.$$

The last argument of min in the last inequality is a result of $P(y'|x) + P(y|x) = 1$, followed by $1 - P(y|x') = P(y'|x')$.

11.2.4 P(benefit) From Admissible Set and Monotonicity

If covariate set $\mathbf{Z}$ satisfies the back-door criterion and Y is monotonic relative to X, then Equation (4.16) can be derived as follows:

$$\begin{aligned} P(\text{benefit}) &= \sum_{\mathbf{z}} P(y|x,\mathbf{z}) \cdot P(\mathbf{z}) - \sum_{\mathbf{z}} P(y|x',\mathbf{z}) \cdot P(\mathbf{z})) \\ &= \sum_{\mathbf{z}} [P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(\mathbf{z}) \\ &= \mathbb{E}_{\mathbf{z}}[P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})]. \end{aligned}$$

11.2.5 Z-Stratified PN from Admissible Set and Monotonicity

If covariate set $\mathbf{Z}$ satisfies the back-door criterion and Y is monotonic relative to X, then Equation (4.21) can be derived as follows:

$$\begin{aligned} \text{PN}_{\mathbf{z}} &= \sum_{\mathbf{z}} \frac{P(y|\mathbf{z}) - P(y|x',\mathbf{z})}{P(x,y|\mathbf{z})} \cdot P(\mathbf{z}|x,y) \\ &= \sum_{\mathbf{z}} \frac{P(y|\mathbf{z}) - P(y|x',\mathbf{z})}{P(x,y)} \cdot P(\mathbf{z}) \qquad (11.4) \\ &= \mathbb{E}_{\mathbf{z}}[P(y|\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(x,y)^{-1}. \end{aligned}$$

The simplification in (11.4) follows from:

$$\frac{P(\mathbf{z}|x,y)}{P(x,y|\mathbf{z})} = \frac{\frac{P(x,y,\mathbf{z})}{P(x,y)}}{\frac{P(x,y,\mathbf{z})}{P(\mathbf{z})}} = \frac{P(\mathbf{z})}{P(x,y)}.$$

11.2.6 Z-Stratified P(benefit) from Admissible Set and Monotonicity

If covariate set $\mathbf{Z}$ satisfies the back-door criterion and Y is monotonic relative to X, then Equation (4.22) can be derived as follows:

$$\begin{aligned} P(\text{benefit})_{\mathbf{z}} &= \sum_{\mathbf{z}} P(y|x,\mathbf{z}) \cdot P(\mathbf{z}) - \sum_{\mathbf{z}} P(y|x',\mathbf{z}) \cdot P(\mathbf{z})) \\ &= \sum_{\mathbf{z}} [P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(\mathbf{z}) \\ &= \mathbb{E}_{\mathbf{z}}[P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})]. \end{aligned}$$

11.2.7 Z-Stratified PN from Admissible Set Bounds

If covariate set $\mathbf{Z}$ satisfies the back-door criterion, then the lower bound in Equation (4.24) can be derived as follows:

$$\begin{aligned} \text{PN}_{\mathbf{z}} &\geqslant \sum_{\mathbf{z}} \text{PN}_{\text{lower-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\ &= \sum_{\mathbf{z}} \max\left\{0, 1 - \frac{P(y|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \qquad (11.5) \\ &= \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\right\} \cdot \frac{P(\mathbf{z}|x,y)}{P(y|x,\mathbf{z})} \\ &= P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \qquad (11.6) \end{aligned}$$

If covariate set $\mathbf{Z}$ satisfies the back-door criterion, then the upper bound in Equation

(4.25) can be derived as follows:

$$
\begin{aligned}
\mathrm{PN}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \mathrm{PN}_{\text{upper-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\
&= \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \\
&= 1 - \left(1 - \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \\
&= 1 - \left(1 + \sum_{\mathbf{z}} \max\left\{-1, -\frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \qquad (11.7) \\
&= 1 - \sum_{\mathbf{z}} \left(P(\mathbf{z}|x,y) + \max\left\{-1, -\frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \\
&= 1 - \sum_{\mathbf{z}} \max\left\{1-1, 1 - \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \\
&= 1 - P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y'|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \qquad (11.8)
\end{aligned}
$$

The simplifications in (11.6) and (11.8) follow from

$$
\frac{P(\mathbf{z}|x,y)}{P(y|x,\mathbf{z})} = \frac{\frac{P(y,\mathbf{z}|x)}{P(y|x)}}{\frac{P(y,\mathbf{z}|x)}{P(\mathbf{z}|x)}} = \frac{P(\mathbf{z}|x)}{P(y|x)}. \qquad (11.9)
$$

The transition from min to max in (11.7) follows from

$$
-\min\{a,b\} = \max\{-a,-b\}.
$$

11.2.8 P(**double-harmed**) **with Pure Mediator Lower Bound**

Equation (4.47) is derived as follows:

$$\begin{aligned}
P(\text{double-harmed}) &= P(y'_m, y_{m'}, m'_x, m_{x'}) \\
&= P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}) \\
&\geqslant \max\{0, P(y'_m) - P(y'_{m'})\} \cdot \max\{0, P(m'_x) - P(m'_{x'})\} \\
&= \max\{0, P(y_{m'}) - P(y_m)\} \cdot \max\{0, P(m_{x'}) - P(m_x)\} \\
&= \max\{0, [P(y_{m'}) - P(y_m)] \cdot [P(m_{x'}) - P(m_x)]\} \\
&= \max\{0, [P(y_m) - P(y_{m'})] \cdot [P(m_x) - P(m_{x'})]\}.
\end{aligned}$$

11.2.9 P(**double-harmed**) **with Pure Mediator Upper Bound**

Equation (4.48) is derived as follows:

$$\begin{aligned}
P(\text{double-harmed}) &= P(y'_m, y_{m'}, m'_x, m_{x'}) \\
&= P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}) \\
&\leqslant \min\{P(y_{m'}), P(y'_m)\} \cdot \min\{P(m_{x'}), P(m'_x)\} \\
&= \begin{cases}
P(y_{m'}) \cdot P(m_{x'}), & P(y_{m'}) \leqslant P(y'_m) \wedge P(m_{x'}) \leqslant P(m'_x), \\
P(y_{m'}) \cdot P(m'_x), & P(y_{m'}) \leqslant P(y'_m) \wedge P(m_{x'}) \geqslant P(m'_x), \\
P(y'_m) \cdot P(m_{x'}), & P(y_{m'}) \geqslant P(y'_m) \wedge P(m_{x'}) \leqslant P(m'_x), \\
P(y'_m) \cdot P(m'_x), & P(y_{m'}) \geqslant P(y'_m) \wedge P(m_{x'}) \geqslant P(m'_x).
\end{cases} \\
&= \begin{cases}
P(y_{m'}) \cdot P(m_{x'}), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\
P(y_{m'}) \cdot P(m'_x), & P(y_m) \leqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}), \\
P(y'_m) \cdot P(m_{x'}), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \leqslant P(m'_{x'}), \\
P(y'_m) \cdot P(m'_x), & P(y_m) \geqslant P(y'_{m'}) \wedge P(m_x) \geqslant P(m'_{x'}).
\end{cases}
\end{aligned}$$

11.3 Appendix for Chapter 9

11.3.1 Bounds on P(benefit) with Non-Binary Ordinal Outcomes

The following lower and upper bound proofs provide conceptual and mathematical proofs of the bounds on P(benefit) with non-binary ordinal outcomes.

11.3.1.1 Lower Bound

The first arguments to the outer max and min of Equations (9.5) and (9.6) and their corresponding expressions in Equation (9.4) all come from Fréchet Inequalities (Section 3.17), where $A = y_{i_{x_2}}$ and $B = y_{j_{x_1}}$.

Equation (9.4) can now serve as a base for deriving tight bounds on the overall P(benefit). For the third argument to the max function, $P(y_i) - P(y_{i_{x_1}})$, we can imagine the individuals represented by y_i and the individuals represented by $y_{i_{x_1}}$. $Y = y_i$ (outcome y_i in an observational study) consists of some individuals who benefit from treatment (the ones who chose treatment and had outcome y_i but would have had outcome y_j where $j < i$ if they chose no treatment, and the ones who chose no treatment and had outcome y_i but would have had outcome y_k where $k > i$ if they chose treatment), some individuals who are harmed by treatment (the ones who chose no treatment and had outcome y_i but would have had outcome y_j where $j < i$, and the ones who chose treatment and had outcome y_i but would have had outcome y_k where $k > i$ if they chose no treatment), and all individuals who have outcome y_i regardless of treatment (the unaffected). $Y_{x_1} = y_i$ (outcome y_i had no treatment been administered) consists of all individuals who benefit from treatment and would have had outcome y_k where $k > i$ had they been administered treatment $(y_{i_{x_1}}, y_{k_{x_2}})$, all individuals who are harmed by treatment and would have had outcome y_j where $j < i$ had they been administered treatment $(y_{i_{x_1}}, y_{j_{x_2}})$, and all individuals who have outcome y_i regardless

of treatment (the immune, $y_{i_{x_1}}, y_{i_{x_2}}$). Therefore,

$$\begin{aligned}
P(y_i) &= P(\text{benefiters avoiding treatment, would've had outcome } y_k \\
&\quad + \text{benefiters choosing treatment, would've had outcome } y_j \\
&\quad + \text{harmed avoiding treatment, would've had outcome } y_j \\
&\quad + \text{harmed choosing treatment, would've had outcome } y_k \\
&\quad + \text{all immune with outcome } y_i), \\
P(y_i) &= P(y_{>i_{x_2}}, y_{i_{x_1}}|x_1) \cdot P(x_1) + P(y_{i_{x_2}}, y_{<i_{x_1}}|x_2) \cdot P(x_2) \\
&\quad + P(y_{<i_{x_2}}, y_{i_{x_1}}|x_1) \cdot P(x_1) + P(y_{i_{x_2}}, y_{>i_{x_1}}|x_2) \cdot P(x_2) \\
&\quad + P(y_{i_{x_2}}, y_{i_{x_1}}, x_1) + P(y_{i_{x_2}}, y_{i_{x_1}}, x_2), \\
P(y_{i_{x_1}}) &= P(\text{all benefiters who would've had outcome } y_k \text{ if treated} \\
&\quad + \text{all harmed who would've had outcome } y_j \text{ if treated} \\
&\quad + \text{all immune with outcome } y_i), \\
P(y_{i_{x_1}}) &= P(y_{>i_{x_2}}, y_{i_{x_1}}) + P(y_{<i_{x_2}}, y_{i_{x_1}}) + P(y_{i_{x_2}}, y_{i_{x_1}}), \\
P(y_i) - P(y_{i_{x_1}}) &= P(y_{<>i_{x_2}}, y_{i_{x_1}}, x_1) - P(y_{<>i_{x_2}}, y_{i_{x_1}}) + P(y_{i_{x_2}}, y_{<>i_{x_1}}, x_2) \\
&= P(y_{i_{x_2}}, y_{<>i_{x_1}}, x_2) - P(y_{<>i_{x_2}}, y_{i_{x_1}}, x_2), \\
\sum_{i=2}^{n} \left[P(y_i) - P(y_{i_{x_1}})\right] &= P(y_{2_{x_2}}, y_{<>2_{x_1}}, x_2) - P(y_{<>2_{x_2}}, y_{2_{x_1}}, x_2) \\
&\quad + P(y_{3_{x_2}}, y_{<>3_{x_1}}, x_2) - P(y_{<>3_{x_2}}, y_{3_{x_1}}, x_2) \\
&\quad + \quad \vdots \quad - \quad \vdots \\
&\quad + P(y_{n_{x_2}}, y_{<n_{x_1}}, x_2) - P(y_{<n_{x_2}}, y_{n_{x_1}}, x_2), \qquad (11.10)
\end{aligned}$$

where $j < i < k$; $y_{<m}$ means any y_i such that $i < m$; $y_{>m}$ means any y_i such that $i > m$; and $y_{<>m}$ means any y_i such that $i < m$ or $i > m$.

Two lower bounds on P(benefit) can be obtained from Equation (11.10). First, the last row consists of all benefiting units that have outcome $Y = y_n$ and were treated with $X = x_2$ subtracted by all harmed units who would have had outcome $Y = y_n$ had they been treated

with $X = x_1$ but were treated with $X = x_2$. This doesn't include all benefiters and subtracts some quantity of non-benefiters, but it is sufficient to be a lower bound.

Second, observe that the harmed units, $(y_{m_{x_2}}, y_{>m_{x_1}})$, get added in the $(m-1)$th row of left column and subsequently subtracted in the entire right column starting at the mth row. We end up with some benefiters, but not all, and a negative proportion of individuals harmed according to $(y_{1_{x_2}}, y_{2_{x_1}})$. This is sufficient to be a lower bound. We can further simplify this expression:

$$\begin{aligned}\sum_{i=2}^{n} \left[P(y_i) - P(y_{i_{x_1}})\right] &= \sum_{i=2}^{n} P(y_i) - \sum_{i=2}^{n} P(y_{i_{x_1}}) \\ &= 1 - P(y_1) - \left[1 - P(y_{1_{x_1}})\right] \\ &= P(y_{1_{x_1}}) - P(y_1).\end{aligned}$$

Similarly, for the fourth argument to the max function of Equation (9.4),

$$
\begin{aligned}
P(y_{ix_2}) &= P(y_{ix_2}, y_{>ix_1}) + P(y_{ix_2}, y_{<ix_1}) \\
&\quad + P(y_{ix_2}, y_{ix_1}), \\
P(y_i) &= P(y_{>ix_2}, y_{ix_1}|x_1) \cdot P(x_1) + P(y_{ix_2}, y_{<ix_1}|x_2) \cdot P(x_2) \\
&\quad + P(y_{<ix_2}, y_{ix_1}|x_1) \cdot P(x_1) + P(y_{ix_2}, y_{>ix_1}|x_2) \cdot P(x_2) \\
&\quad + P(y_{ix_2}, y_{ix_1}, x_1) + P(y_{ix_2}, y_{ix_1}, x_2), \\
P(y_{ix_2}) - P(y_i) &= P(y_{ix_2}, y_{<>ix_1}, x_1) - P(y_{<>ix_2}, y_{ix_1}, x_1), \\
\sum_{i=2}^{n} \left[P(y_{ix_2}) - P(y_i)\right] &= P(y_{2x_2}, y_{<>2x_1}, x_1) - P(y_{<>2x_2}, y_{2x_1}, x_1) \\
&\quad + P(y_{3x_2}, y_{<>3x_1}, x_1) - P(y_{<>3x_2}, y_{3x_1}, x_1) \\
&\quad + \quad \vdots \quad - \quad \vdots \\
&\quad + P(y_{nx_2}, y_{<nx_1}, x_1) - P(y_{<nx_2}, y_{nx_1}, x_1) \\
&= \sum_{i=2}^{n} P(y_{ix_2}) - \sum_{i=2}^{n} P(y_i) \\
&= 1 - P(y_{1x_2}) - [1 - P(y_1)] \\
&= P(y_1) - P(y_{1x_2}).
\end{aligned}
$$

The final term in the non-binary ordinal outcome $P(\text{benefit})$'s lower bound is unrelated to any terms in the Tian-Pearl lower bound. Following a similar type of strategy in [LDD18],

a potential lower bound can be derived in the following way:

$$
\begin{aligned}
& \forall k \in [n] \\
P(\text{benefit}) &= \sum_{\substack{1 \leqslant i,j \leqslant n \\ \text{s.t. } i>j}} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) \\
&\geqslant \sum_{j=1}^{k} \sum_{i=k+1}^{n} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) - \sum_{j=1}^{k} \sum_{i=1}^{k} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) \\
&\geqslant \sum_{j=1}^{k} \sum_{i=1}^{n} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) - \sum_{j=1}^{n} \sum_{i=1}^{k} P\left(y_{i_{x_2}}, y_{j_{x_1}}\right) \\
&= \sum_{j=1}^{k} P\left(y_{j_{x_1}}\right) - \sum_{i=1}^{k} P\left(y_{i_{x_2}}\right) \\
&= \sum_{j=1}^{k} P\left(y_{j_{x_1}}\right) - \sum_{j=1}^{k} P\left(y_{j_{x_2}}\right).
\end{aligned}
$$

Since k can be any value between 1 and n,

$$
P(\text{benefit}) \geqslant \max_{1 \leqslant i < n} \sum_{j=1}^{i} \left[P(y_{j_{x_1}}) - P(y_{j_{x_2}}) \right]. \tag{11.11}
$$

Note the max function in Equation (11.11) ranges over all indices of y_i except y_n. This is because the entire expression for the lower bound would equate to $1 - 1 = 0$, which the Fréchet Inequality for the lower bound already covers.

11.3.1.2 Upper Bound

Let us continue to use Equation (9.4) as a base for deriving a tight upper bound on $P(\text{benefit})$. Just like the lower bound, the first two arguments to min come from the right-side Fréchet Inequality in Eq. (3.17).

The third argument evolves into an upper bound according to the following derivation.

$$
\begin{aligned}
P(x_2, y_i) &= P(\text{some treated benefiters, untreated outcome is } y_{<i} \\
&\quad + \text{some treated harmed, untreated outcome is } y_{>i} \\
&\quad + \text{all immune with outcome } y_i) \\
&= P(y_{i_{x_2}}, y_{<i_{x_1}}, x_2) + P(y_{i_{x_2}}, y_{>i_{x_1}}, x_2) \\
&\quad + P(y_{i_{x_2}}, y_{i_{x_1}}, x_2), \\
P(x_1, y_j) &= P(\text{some untreated benefiters, treated outcome is } y_{>j} \\
&\quad + \text{some untreated harmed, treated outcome is } y_{<j} \\
&\quad + \text{all immune with outcome } y_j) \\
&= P(y_{>j_{x_2}}, y_{j_{x_1}}, x_1) + P(y_{<j_{x_2}}, y_{j_{x_1}}, x_1) \\
&\quad + P(y_{j_{x_2}}, y_{j_{x_1}}, x_1), \\
P(x_2, y_i) + P(x_1, y_j) &= P(y_{i_{x_2}}, y_{<>i_{x_1}}, x_2) + P(y_{<>j_{x_2}}, y_{j_{x_1}}, x_1) \\
&\quad + P(y_{i_{x_2}}, y_{i_{x_1}}, x_2) + P(y_{j_{x_2}}, y_{j_{x_1}}, x_1), \\
\sum_{i=2}^{n} P(x_2, y_i) + \sum_{j=1}^{n-1} P(x_1, y_j) &= P(\text{all treated and untreated benefiters} \\
&\quad + \text{all treated harmed except } y_{1_{x_2}} \\
&\quad + \text{all untreated harmed except } y_{n_{x_1}} \\
&\quad + \text{all treated immune except } y_{1_{x_2}}, y_{1_{x_1}} \\
&\quad + \text{all untreated immune except } y_{n_{x_2}}, y_{n_{x_1}}) \qquad (11.12) \\
&= P(x_2) - P(x_2, y_1) + P(x_1) - P(x_1, y_n) \\
&= 1 - P(x_2, y_1) - P(x_1, y_n),
\end{aligned}
$$

where $j < i < k$; $y < m$ means any y_i such that $i < m$; $y > m$ means any y_i such that $i > m$; and $y <> m$ means any y_i such that $i < m$ or $i > m$.

Let us rewrite the fourth argument to the min function of Equation (9.4):

$$\begin{aligned} P\left(y_{ix_2}, y_{jx_1}\right) &\leqslant P(y_{ix_2}) - P(y_{ix_1}) + P(x_2, y_j) + P(x_1, y_i) \\ &= P(y_{ix_2}) - [1 - P(y_{<>ix_1})] + (1 - [P(x_2, y_i) + P(x_1, y_j)]) \\ &= P(y_{ix_2}) + P(y_{<>ix_1}) - [P(x_2, y_i) + P(x_1, y_j)]. \end{aligned}$$

Now we can use the fact that $\sum_{i=2}^{n} y_{ix_2}$ represents all benefiting individuals, all harmed individuals except y_{1x_2}, and all immune individuals except (y_{1x_2}, y_{1x_1}), while $y_{<>nx_1} = y_{<nx_1}$ represents all benefiters, all harmed except y_{nx_1}, and all immune except (y_{nx_2}, y_{nx_1}). Com-

bining that with Equation (11.12),

$$\begin{aligned}
&\sum_{i=2}^{n} P(y_{i_{x_2}}) + \sum_{i=1}^{n-1} P(y_{i_{x_1}}) - \left[\sum_{i=2}^{n} P(x_2, y_i) + \sum_{j=1}^{n-1} P(x_1, y_j)\right] \\
&\qquad = \sum_{i=2}^{n} \left[P(y_{i_{x_2}}) - P(x_2, y_i)\right] + \sum_{i=1}^{n-1} \left[P(y_{i_{x_1}}) - P(x_1, y_j)\right] \\
&\qquad = P(\text{all benefiters} + \text{all harmed except } y_{1_{x_2}} \\
&\qquad\quad + \text{all immune except } y_{1_{x_2}}, y_{1_{x_1}}) \\
&\qquad\quad + P(\text{all benefiters} + \text{all harmed except } y_{n_{x_1}} \\
&\qquad\qquad + \text{all immune except } (y_{n_{x_2}}, y_{n_{x_1}})) \\
&\qquad\quad - P(\text{all benefiters} \\
&\qquad\qquad + \text{all treated harmed except } y_{1_{x_2}} \\
&\qquad\qquad + \text{all untreated harmed except } y_{n_{x_1}} \\
&\qquad\qquad + \text{all treated immune except } (y_{1_{x_2}}, y_{1_{x_1}}) \\
&\qquad\qquad + \text{all untreated immune except } (y_{n_{x_2}}, y_{n_{x_1}})) \\
&\qquad = P(\text{benefit}) + P(\text{untreated harmed except } y_{1_{x_2}}) \\
&\qquad\quad + P(\text{treated harmed except } y_{n_{x_1}}) \\
&\qquad\quad + P(\text{untreated immune except } (y_{1_{x_2}}, y_{1_{x_1}})) \\
&\qquad\quad + P(\text{treated immune except } (y_{n_{x_2}}, y_{n_{x_1}})) \\
&\qquad \leqslant P(\text{benefit}).
\end{aligned}$$

Let us simplify:

$$\begin{aligned}
&\sum_{i=2}^{n} \left[P(y_{i_{x_2}}) - P(x_2, y_i)\right] + \sum_{i=1}^{n-1} \left[P(y_{i_{x_1}}) - P(x_1, y_j)\right] \\
&\qquad = 1 - P(y_{1_{x_2}}) - [P(x_2) - P(x_2, y_1)] + 1 - P(y_{n_{x_1}}) - [P(x_1) - P(x_1, y_n)] \\
&\qquad = 2 - P(x_2) - P(x_1) - P(y_{1_{x_2}}) - P(y_{n_{x_1}}) + P(x_2, y_1) + P(x_1, y_n) \\
&\qquad = 1 - P(y_{1_{x_2}}) - P(y_{n_{x_1}}) + P(x_2, y_1) + P(x_1, y_n).
\end{aligned}$$

The final term in the non-binary ordinal outcome $P(\text{benefit})$'s upper bound is unrelated to any terms in the Tian-Pearl upper bound. Following the same strategy as the derivation of Equation (11.11), a potential upper bound can be derived in the following way:

$$
\begin{aligned}
&\forall k \in [n] \\
P(\text{benefit}) &= \sum_{\substack{1 \leqslant i,j \leqslant n \\ \text{s.t. } i > j}} P\left(y_{ix_2}, y_{jx_1}\right) \\
&\leqslant 1 - \sum_{i=1}^{k} \sum_{j=k}^{n} P\left(y_{ix_2}, y_{jx_1}\right) \\
&= 1 - \left[\sum_{i=1}^{k} \sum_{j=1}^{n} P\left(y_{ix_2}, y_{jx_1}\right) - \sum_{i=1}^{k} \sum_{j=1}^{k-1} P\left(y_{ix_2}, y_{jx_1}\right) \right] \\
&\leqslant 1 - \left[\sum_{i=1}^{k} \sum_{j=1}^{n} P\left(y_{ix_2}, y_{jx_1}\right) - \sum_{i=1}^{n} \sum_{j=1}^{k-1} P\left(y_{ix_2}, y_{jx_1}\right) \right] \\
&= 1 - \left[\sum_{i=1}^{k} P\left(y_{ix_2}\right) - \sum_{j=1}^{k-1} P\left(y_{jx_1}\right) \right] \\
&= 1 - \left[\sum_{j=1}^{k} P\left(y_{jx_2}\right) - \sum_{j=1}^{k} P\left(y_{jx_1}\right) + P\left(y_{kx_1}\right) \right] \\
&= 1 - \left[\sum_{i=1}^{k} P\left(y_{ix_2}\right) - \sum_{j=1}^{k} P\left(y_{jx_1}\right) + P\left(y_{kx_1}\right) \right].
\end{aligned}
$$

Since k can be any value between 1 and n,

$$
P(\text{benefit}) \leqslant 1 - \max_{1 \leqslant i \leqslant n} \left\{ \sum_{j=1}^{i} \left[P(y_{jx_2}) - P(y_{jx_1}) \right] + P\left(y_{ix_1}\right) \right\}.
$$

11.3.2 Binary Outcome $P(\textbf{benefit})$ under MITE

Theorem 9.6.2 is applied to $P(\text{benefit})$:

$$\begin{aligned} P(\text{benefit}) &= P(y_{2_{x_2}}, y_{1_{x_1}}) \\ &= \sum_{k=1}^{1} \left[P(y_{k_{x_1}}) - P(y_{k_{x_2}}) \right] \\ &= P(y_{1_{x_1}}) - P(y_{1_{x_2}}) \\ &= \left[1 - P(y_{2_{x_1}}) \right] - \left[1 - P(y_{2_{x_2}}) \right] \\ &= P(y_{2_{x_2}}) - P(y_{2_{x_1}}). \end{aligned}$$

11.3.3 Binary Outcome $P(\textbf{immunity})$ under MITE

Theorem 9.6.2 is applied to $P(\text{immunity})$:

$$\begin{aligned} P(\text{immunity}) &= P(y_{2_{x_2}}, y_{2_{x_1}}) \\ &= \sum_{k=1}^{2} P(y_{k_{x_2}}) - \sum_{k=1}^{1} P(y_{k_{x_1}}) \\ &= P(y_{1_{x_2}}) + P(y_{2_{x_2}}) - P(y_{1_{x_1}}) \\ &= 1 - P(y_{1_{x_1}}) \qquad (11.13) \\ &= P(y_{2_{x_1}}). \end{aligned}$$

Equality (11.13) is due to the second probability axiom: $P(y_{1_{x_2}}) + P(y_{2_{x_2}}) = 1$.

11.3.4 Binary Outcome $P(\textbf{doom})$ under MITE

Theorem 9.6.2 is applied to $P(\text{doom})$:

$$\begin{aligned} P(\text{doom}) &= P(y_{1_{x_2}}, y_{1_{x_1}}) \\ &= \sum_{k=1}^{1} P(y_{1_{x_2}}) - \sum_{k=1}^{0} P(y_{1_{x_1}}) \\ &= P(y_{1_{x_2}}). \end{aligned}$$