TECHNICAL REPORT
R-506
June 2021

UNIVERSITY OF CALIFORNIA
Los Angeles

Estimating Individualized Causes of Effects by Leveraging Population Data

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Scott Allen Mueller

2021

ABSTRACT OF THE THESIS

Estimating Individualized Causes of Effects
by Leveraging Population Data

by

Scott Allen Mueller
Master of Science in Computer Science
University of California, Los Angeles, 2021
Professor Judea Pearl, Chair

Most analyses in the past three decades concerned 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. For example, personalized medicine concerns the probability of a drug being the *cause* of survival: resulting in a favorable outcome if taken *and* unfavorable if avoided. One reason for this imbalance is that tools for estimating the probability of causation from data require counterfactual logic. Bounds on these probabilities are often too loose to be informative and the assumptions necessary for point estimates are often too strong to be defensible. The objective of this thesis is to develop and test techniques for achieving narrower bounds on the probabilities of causation, with minimal assumptions. These more accurate estimates are achieved by incorporating a causal model and covariate data.

The thesis of Scott Allen Mueller is approved.

Adnan Darwiche

Chad Hazlett

Judea Pearl, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Six sections within chapters 3 and 4 build upon and extend work done by Ang Li and myself. That work was written in [MLP21]. In particular, sections 3.2.2 and 3.3.2 advanced this work with conditions formalized under which the new PNS bounds have superiority over the old PNS bounds, practical usage guidelines, and additional analyses, examples, and graphical criterion. Sections 4.1.1 and 4.2.1 similarly advanced this work with a further developed derivation on PNS bounds with mediators, relaxed graphical criterion that is more flexible, and graphical methods to test applicability of formulas. Sections 3.4 and 3.5 are derived directly from our paper.

This research was supported in parts by grants from the National Science Foundation [#IIS1704932], Office of Naval Research [#N00014-17-S-12091 and #N00014-21-1-2351] and Toyota Research Institute of North America [#PO-000897].

CHAPTER 1

Introduction

Most analyses in the past three decades concerned 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]. For example, personalized medicine concerns the probability of a drug being the *cause* of survival: resulting in a favorable outcome if taken *and* unfavorable if avoided.

One reason for this imbalance is that CoE requires understanding conflicting, or counterfactual, probabilities at the individual level. Observing survival when treated and death when untreated hits at the heart of the fundamental problem of causal inference: we can only observe a single treatment and outcome in an individual. For this reason, without strong assumptions or knowledge of the underlying functional form, we can generally only obtain bounds, as opposed to point estimates, on probabilities of causation from statistical data. Unfortunately, bounds on these probabilities 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 analysis in (Mueller and Pearl, 2020) focused on gender. However, [MLP21] used the following, more realistic, scenario of a machine learning algorithm discovering a high correlation between age and recovery. We classify ages into two groups: under sixty years old and over 59 years old. From an RCT, it is determined that older people have an average causal effect (ACE) 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 ACE among younger people to be 10%, as 55% survive when vaccinated and 45% survive when unvaccinated. If just comparing ACEs, it would seem that the vaccine is $20\% - 10\% = 10\%$ 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. The proportion of individuals who would benefit from treatment is known as the probability of necessity and sufficiency (PNS). This quantity is the probability of an individual surviving if vaccinated and dying if unvaccinated, precisely what we're looking for. We can then compare the PNS for elders versus young. Traditional counterfactual analysis [TP00] yields bounds on the PNS 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 the PNS among the younger group

is calculated to be between 10% and 55%. This is also a large range and it significantly overlaps with the PNS bounds among the older subpopulation. Which group should receive priority for vaccination?

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 chapter 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 PNS, 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 thesis explores methods to compute narrower bounds on popular probabilities of causation. Beyond causal effects, it's 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. Chapter 3 will demonstrate this with covariates even when they're not needed for identification of causal effects.

The next chapters are organized as follows. Chapter 2 offers descriptions and existing analyses of the three most prominent probabilities of causation. Chapter 3 covers narrowing these three probabilities of causation using covariate data, including formulas, proofs, and graphical criterion. Chapter 4 provides the same type of analyses for mediators. Chapter 5 gives tools to combine multiple covariate and mediator data with more complicated graphs to further narrow bounds. Finally, chapter 6 concludes with a discussion of the results and future directions.

CHAPTER 2

Probabilities of Causation

This chapter reviews three important probabilities of causation as defined in [Pea99]: the probability of necessity (PN), the probability of sufficiency (PS), and the probability of necessity and sufficiency (PNS). Causal diagrams [Pea95, SGS00, Pea09, KF09] and the language of counterfactuals in its structural model semantics, as given in [BP13, GP98, Hal00] are used in this and following chapters.

2.1 Notation

The following notational conventions will be adopted for the remainder of this thesis. 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: $P(y_x)$, $P(y_{x'})$, $P(y'_x)$, and $P(y'_{x'})$.

2.2 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 2.2.1 (Probability of Necessity (PN))**.** *Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X = true$ and $Y = true$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression [Pea99]*

$$\begin{aligned} \text{PN} &\triangleq P(Y_{x'} = false | X = true, Y = true) \\ &\triangleq P(y'_{x'} | x, y) \end{aligned} \tag{2.1}$$

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.

2.3 Probability of Sufficiency

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 2.3.1 (Probability of Sufficiency (PS))**.** *[Pea99]*

$$\text{PS} \triangleq P(y_x|y', x') \tag{2.2}$$

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].

2.4 Probability of Necessity and Sufficiency

Many labs are working on treatments for this disease. Imagine a randomized controlled trial (RCT) for a new treatment 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. This is known as the Probability of Necessity and Sufficiency (PNS). This scenario will be examined a little further in section 2.6. Clearly, from a public policy viewpoint a drug that kills some patients and cures others should be treated with greater caution than one that has no effect whatsoever on any individual.

Definition 2.4.1 (Probability of Necessity and Sufficiency (PNS))**.** *[Pea99]*

$$\text{PNS} \triangleq P(y_x, y'_{x'}) \tag{2.3}$$

This is the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

2.5 Bounds

The three probabilities mentioned above, PN, PS, and PNS, 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. However, informative bounds for PN and PNS can be derived when experimental and observational data are available. 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 programming (LP) problem describing these probabilities of causation. Li and Pearl [LP19] provide a theoretical proof of the tight bounds for PN, PS, PNS, 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

thesis:

$$\max\left\{0, \frac{P(y) - P(y_{x'})}{P(x,y)}\right\} \leqslant \text{PN} \leqslant \min\left\{1, \frac{P(y'_{x'}) - P(x',y')}{P(x,y)}\right\}, \tag{2.4}$$

$$\max\left\{0, \frac{P(y_x) - P(y)}{P(x',y')}\right\} \leqslant \text{PS} \leqslant \min\left\{1, \frac{P(y_x) - P(x,y)}{P(x',y')}\right\}, \tag{2.5}$$

$$\max\left\{\begin{array}{l} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array}\right\} \leqslant \text{PNS} \leqslant \min\left\{\begin{array}{l} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + P(x',y) \; + P(x,y') \end{array}\right\}. \tag{2.6}$$

Bounds for a specific subpopulation, defined by a set C of pretreatment characteristics, usually yield narrower bounds because variation among units is reduced in the subpopulation compared with the overall population. These subpopulation bounds can be obtained by simply conditioning each probability in the bounds above on $C = c$.

2.5.1 Exogeneity

Definition 2.5.1 (Exogeneity)**.** *A variable X is said to be exogenous for Y in model M iff [TP00]*

$$Y_x \perp\!\!\!\perp X \quad \textit{and} \quad Y_{x'} \perp\!\!\!\perp X. \tag{2.7}$$

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 PNS become[1]:

$$\frac{\max\{0, P(y|x) - P(y|x')\}}{P(y|x)} \leqslant \text{PN} \leqslant \frac{\min\{P(y|x), P(y'|x')\}}{P(y|x)}, \tag{2.8}$$

$$\frac{\max\{0, P(y|x) - P(y|x')\}}{P(y'|x')} \leqslant \text{PS} \leqslant \frac{\min\{P(y|x), P(y'|x')\}}{P(y'|x')}, \tag{2.9}$$

$$\max\{0, P(y|x) - P(y|x')\} \leqslant \text{PNS} \leqslant \min\{P(y|x), P(y'|x')\}. \tag{2.10}$$

Conceptually, causal models should be informative for PN, PS, and PNS 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 PN and PNS.

Chapters 3, 4, and 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 the 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].

2.6 Toy Example

Continuing from PNS example in section 2.4 where an RCT was conducted with 40% cured and 60% not cured in both treatment and control groups, is it possible that the treatment actually cured 40% of patients? The ACE is null, but the Tian-Pearl bounds can be calculated to see what the possible PNS probabilities are. Observational data is unavailable, so

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

only the first two arguments to min and max will be used:

$$\max\{0, P(y_x) - P(y_{x'})\} \leqslant \text{PNS} \leqslant \min\{P(y_x), P(y'_{x'})\}$$
$$\max\{0, 0.4 - 0.4\} \leqslant \text{PNS} \leqslant \min\{0.4, 0.6)\}$$
$$0 \leqslant \text{PNS} \leqslant 0.4.$$

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 ACE 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 [PM21] 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.

2.7 Identification

Monotonicity is the condition that treatment never harms patients:

Definition 2.7.1 (Monotonicity)**.** *A variable Y is said to be monotonic relative to variable X in a causal model M iff [TP00]*

$$y'_x \wedge y_{x'} = \mathit{false}. \tag{2.11}$$

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 this case, point estimates, rather than bounds, can be found for PN, PS, and PNS:

$$\text{PN} = \frac{P(y) - P(y_{x'})}{P(x, y)}, \tag{2.12}$$

$$\text{PS} = \frac{P(y_x) - P(y)}{P(x', y')}, \tag{2.13}$$

$$\text{PNS} = P(y_x) - P(y_{x'}). \tag{2.14}$$

The intuition behind this is best exemplified in the PNS estimate. First, let us split up the population into four groups by how units or individuals respond to treatment. This is displayed in table 2.1.

Response Type	Notation	Counterfactual Nature
Complier	$y_x, y'_{x'}$	Positive outcome if treated, negative outcome if untreated
Always-taker	$y_x, y_{x'}$	Positive outcome regardless of treatment
Never-taker	$y'_x, y'_{x'}$	Negative outcome regardless of treatment
Defier	$y'_x, y_{x'}$	Negative outcome if treated, positive outcome if untreated

Table 2.1: Four response types of units/individuals

The names of these four groups come from instrumental variable (IV) literature in the context of encouraging treatment. For example, a complier complies with the encouragement and takes the treatment if encouraged and doesn't take the treatment if not encouraged. This nomenclature will be repurposed here for probabilities of causation.

Let us go back to the intuition behind the PNS estimate under the monotonicity assumption. Individuals in the treatment group with a positive outcome could be compliers

or always-takers. 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 complier or always-taker groups. Similarly, $P(y'_x)$ is the proportion of individuals belonging to the defier or always-taker groups.

The ACE is defined as $P(y_x) - P(y_{x'})$. This is the proportion of compliers and always-takers minus the proportion of defiers and always-takers, leaving us with the proportion of compliers minus the proportion of defiers. Notice in table 2.1 that the proportion of compliers is $P(y_x, y'_{x'})$, which is precisely the PNS. Therefore, PNS $= P(y_x) - P(y_{x'}) + P(\text{defier})$. This is why the ACE was only a lower bound for PNS. If monotonicity is assumed, then $P(\text{defier}) = P(y'_x, y_{x'}) = 0$ and we finally arrive at the point estimate PNS $= P(y_x) - P(y_{x'})$.

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, the remainder of this thesis will only consider PN and PNS.

CHAPTER 3

Leveraging Covariate Data

The techniques profiting from covariate data will commence with identification of PN and PNS under monotonicity, as defined in definition 2.7.1. 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 a inadmissible covariate set, especially if both experimental and observational data are available.

3.1 Observational Data Under Monotonicity

Tian and Pearl [TP00] pointed out that the PN, PS, and PNS 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 PNS 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:

$$\begin{aligned}
\text{PN} &= \frac{P(y) - \sum_{\mathbf{z}} P(y|x', \mathbf{z}) \cdot P(\mathbf{z})}{P(x, y)}, \\
\text{PS} &= \frac{\sum_{\mathbf{z}} P(y|x, \mathbf{z}) \cdot P(\mathbf{z}) - P(y)}{P(x', y')}, \\
\text{PNS} &= \sum_{\mathbf{z}} P(y|x, \mathbf{z}) \cdot P(\mathbf{z}) - \sum_{\mathbf{z}} P(y|x', \mathbf{z}) \cdot P(\mathbf{z})) \\
&= \sum_{\mathbf{z}} [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}) \\
&= \mathbb{E}_{\mathbf{z}}[P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})].
\end{aligned}$$

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

$$\text{PN}(\mathbf{z}) = P(y'_{x'}|x, y, \mathbf{z}), \tag{3.1}$$

$$\text{PNS}(\mathbf{z}) = P(y_x, y'_{x'}|\mathbf{z}). \tag{3.2}$$

Under monotonicity, these become:

$$\begin{aligned}
\text{PN}(\mathbf{z}) &= \frac{P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z})}{P(x, y|\mathbf{z})}, \\
\text{PNS}(\mathbf{z}) &= P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}).
\end{aligned}$$

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

$$\text{PN}_{\mathbf{z}} = \sum_{\mathbf{z}} \text{PN}(\mathbf{z}) \cdot P(\mathbf{z}|x, y), \tag{3.3}$$

$$\text{PNS}_{\mathbf{z}} = \sum_{\mathbf{z}} \text{PNS}(\mathbf{z}) \cdot P(\mathbf{z})). \tag{3.4}$$

When $\mathbf{Z}$ satisfies the back-door criterion, PN, PS (through an easy derivation from PN),

and PNS require only observational data:

$$\begin{aligned}
\mathrm{PN}_{\mathbf{z}} &= \sum_{\mathbf{z}} \frac{P(y|\mathbf{z}) - P(y|x', \mathbf{z})}{P(x, y|\mathbf{z})} \cdot P(\mathbf{z}|x, y) \\
&= \sum_{\mathbf{z}} \frac{P(y|\mathbf{z}) - P(y|x', \mathbf{z})}{P(x, y)} \cdot P(\mathbf{z}) \qquad (3.5) \\
&= \mathbb{E}_{\mathbf{z}}[P(y|\mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(x, y)^{-1}, \\
\mathrm{PNS}_{\mathbf{z}} &= \sum_{\mathbf{z}} P(y|x, \mathbf{z}) \cdot P(\mathbf{z}) - \sum_{\mathbf{z}} P(y|x', \mathbf{z}) \cdot P(\mathbf{z})) \\
&= \sum_{\mathbf{z}} [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}) \\
&= \mathbb{E}_{\mathbf{z}}[P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})].
\end{aligned}$$

The simplification in (3.5) follows from:

$$\frac{P(\mathbf{z}|x, y)}{P(x, y|\mathbf{z})} = \frac{\frac{P(x,y,\mathbf{z})}{P(x,y)}}{\frac{P(x,y,\mathbf{z})}{P(\mathbf{z})}} = \frac{P(\mathbf{z})}{P(x, y)}. \qquad (3.6)$$

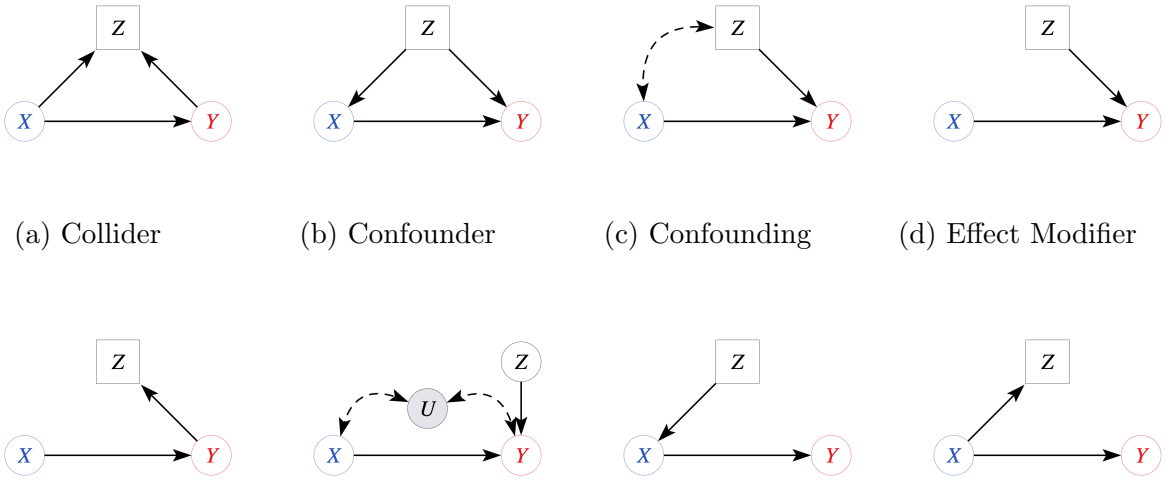
In this way, covariate data has enabled us to identify PN, PS, and PNS in the absence of experimental data. However, monotonicity is a strong and necessary assumption. Kuroki and Cai's stratified PN and PNS will next be used to relax this monotonicity assumption.

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 PNS bounds even in situations where identification is neither feasible nor needed, such as when experimental data are available. The purpose of this thesis 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 PNS. A typical covariate, in the context of the beach going RCT study of sections 2.2 and 2.3, would be measurement of pre-treatment variables in

both treatment and control groups and asking whether it provides a more accurate assessment of PN, PS, and PNS for a typical individual not in the study. Chapter 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 [MLP21, DMM17] for the models described in figure 3.1. 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 3.1: 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 be unmeasurable counterfactual terms. In the rare circumstance that the functional model or structural causal model (SCM) is available, counterfactual terms can be computed. However, an SCM also allows direct computation of the PN, PS, and PNS, 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 3.1h, 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 [CFP20, page 7] point out that conditioning on Z of figure 3.1h reduces variation in X. This can hurt ACE precision in finite samples.

In the case of figures 3.1d, 3.1e, and 3.1f 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 (2.8), (2.9), and (2.10).

Sections 3.2.1 and 3.2.2 assume $\mathbf{Z}$ satisfies the back-door criterion, rendering X exogenous for Y in each stratum $\mathbf{z}$ of $\mathbf{Z}$.

3.2.1 PN Bounds

With the assumption that conditioning on $\mathbf{Z}$ leaves X exogenous, the bounds (2.8) can be applied to PN($\mathbf{z}$) in order to obtain:

$$\max\left\{0, 1 - \frac{P(y|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \leqslant \text{PN}(\mathbf{z}) \leqslant \min\left\{1, \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\}. \tag{3.7}$$

The task is now to bound $\text{PN}_{\mathbf{z}}$, the population PN evaluated with covariate set $\mathbf{Z}$, using the bounds derived at (3.7) for the $\mathbf{z}$-specific PN($\mathbf{z}$). By replacing PN($\mathbf{z}$) in the summation

within equation (3.3) by its lower bound in (3.7), the lower bound for PN is as follows:

$$\begin{aligned}
\mathrm{PN}_{\mathbf{z}} &\geqslant \sum_{\mathbf{z}} \mathrm{PN}_{\text{lower-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\
&= \sum_{\mathbf{z}} \max\left\{0, 1 - \frac{P(y|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \qquad (3.8) \\
&= \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\right\} \cdot \frac{P(\mathbf{z}|x,y)}{P(y|x,\mathbf{z})} \\
&= P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \qquad (3.9)
\end{aligned}$$

Similarly, $\mathrm{PN}(\mathbf{z})$ in the summation within equation (3.3) can be replaced by its upper bound in (3.7) to obtain an upper bound as follows:

$$\begin{aligned}
\mathrm{PN}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \mathrm{PN}_{\text{upper-bound}}(\mathbf{z}) \cdot P(\mathbf{z}|x,y) \\
&= \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \qquad (3.10) \\
&= 1 - \left(1 - \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \\
&= 1 - \left(1 + \sum_{\mathbf{z}} \max\left\{-1, -\frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \qquad (3.11) \\
&= 1 - \sum_{\mathbf{z}} \left(P(\mathbf{z}|x,y) + \max\left\{-1, -\frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y)\right) \\
&= 1 - \sum_{\mathbf{z}} \max\left\{1-1, 1 - \frac{P(y'|x',\mathbf{z})}{P(y|x,\mathbf{z})}\right\} \cdot P(\mathbf{z}|x,y) \\
&= 1 - P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, P(y|x,\mathbf{z}) - P(y'|x',\mathbf{z})\right\} \cdot P(\mathbf{z}|x). \qquad (3.12)
\end{aligned}$$

The simplifications in (3.9) and (3.12) follow from

$$\frac{P(\mathbf{z}|x,y)}{P(y|x,\mathbf{z})} = \frac{\frac{P(y,\mathbf{z}|x)}{P(y|x)}}{\frac{P(y,\mathbf{z}|x)}{P(\mathbf{z}|x)}} = \frac{P(\mathbf{z}|x)}{P(y|x)}. \qquad (3.13)$$

The transition from min to max in (3.11) follows from

$$-\min\{a,b\} = \max\{-a,-b\}.$$

Note that $P(\mathbf{z}|x)$ can be simplified to $P(\mathbf{z})$ in (3.9) and (3.12) if $\mathbf{Z} \perp\!\!\!\perp X$, as in figures 3.1d and 3.1f.

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 3.2.1. *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:*

$$\max\left\{\sum_i a_i, \sum_i b_i\right\} \leqslant \sum_i \max\{a_i, b_i\}. \tag{3.14}$$

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:

$$\min\left\{\sum_i a_i, \sum_i b_i\right\} \geqslant \sum_i \min\{a_i, b_i\}. \tag{3.15}$$

Both (3.14) *and* (3.15) *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 (2.8) with $\text{PN}_{\mathbf{z}}$'s lower bound:

$$\begin{aligned}
\text{PN} &\geqslant \max\left\{0, 1 - \sum_{\mathbf{z}} \frac{P(y|x', \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y)\right\} \\
&= \max\left\{0, \sum_{\mathbf{z}} \frac{P(y|x, \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y) - \sum_{\mathbf{z}} \frac{P(y|x', \mathbf{z})}{P(y|x, \mathbf{z})} P(\mathbf{z}|x, y)\right\} \\
&= \max\left\{0, \sum_{\mathbf{z}} \frac{P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})}{P(y|x)} P(\mathbf{z}|x)\right\} \\
&= P(y|x)^{-1} \max\left\{0, \sum_{\mathbf{z}} [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}|x)\right\}, \qquad (3.16) \\
\text{PN}_{\mathbf{z}} &\geqslant P(y|x)^{-1} \cdot \sum_{\mathbf{z}} \max\left\{0, [P(y|x, \mathbf{z}) - P(y|x', \mathbf{z})] \cdot P(\mathbf{z}|x)\right\} \qquad (3.17)
\end{aligned}$$

The inequality (3.14), 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 (3.16) is inferior to $\text{PN}_{\mathbf{z}}$'s lower bound in (3.17). Let us now compare the Tian-Pearl PN upper bound in (2.8) with $\text{PN}_{\mathbf{z}}$'s upper bound in (3.12):

$$\begin{aligned}
\text{PN} &\leqslant \min\left\{1, \sum_{\mathbf{z}} \frac{P(y'|x', \mathbf{z})}{P(y|x, \mathbf{z})} \cdot P(\mathbf{z}|x, y)\right\} \\
&= \min\left\{\sum_{\mathbf{z}} P(\mathbf{z}|x), \sum_{\mathbf{z}} \frac{P(y'|x', \mathbf{z})}{P(y|x)} \cdot P(\mathbf{z}|x)\right\}, \qquad (3.18) \\
\text{PN}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \min\left\{1, \frac{P(y'|x', \mathbf{z})}{P(y|x)}\right\} \cdot P(\mathbf{z}|x) \\
&= \sum_{\mathbf{z}} \min\left\{P(\mathbf{z}|x), \frac{P(y'|x', \mathbf{z})}{P(y|x)} \cdot P(\mathbf{z}|x)\right\}. \qquad (3.19)
\end{aligned}$$

The inequality (3.15), 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 (3.18) is inferior to $\text{PN}_{\mathbf{z}}$'s lower bound in (3.19).

From lemma 3.2.1, 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 2.12 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 3.2.3 will graphically demonstrate when these bounds can be used. Finally, examples will illustrate the application of covariate-benefiting bounds.

3.2.2 PNS Bounds

Tian and Pearl [TP00] provided bounds on PNS are provided 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:

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

Narrower bounds than this cannot be obtained using the summation of maximums or minimums technique in lemma 3.2.1 because there are not different minimum or maximum options to choose from in each stratum of Z.

With the same conditional ignorability assumption of section 3.2.1, the bounds 2.10 can

be applied to PNS($\mathbf{z}$) in order to obtain:

$$\max\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\} \leqslant \text{PNS}(\mathbf{z}) \leqslant \min\{P(y|x,\mathbf{z}), P(y'|x',\mathbf{z})\}. \tag{3.20}$$

With the same approach taken in section 3.2.1, the $\text{PNS}_{\mathbf{z}}$ lower bound is as follows:

$$\begin{aligned}
\text{PNS}_{\mathbf{z}} &\geqslant \sum_{\mathbf{z}} \text{PNS}_{\text{lower-bound}}(\mathbf{z}) \cdot P(\mathbf{z}) \\
&= \sum_{\mathbf{z}} \max\{0, P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})\} \cdot P(\mathbf{z}) \qquad (3.21) \\
&= \sum_{\mathbf{z}} \max\{0, [P(y|x,\mathbf{z}) - P(y|x',\mathbf{z})] \cdot P(\mathbf{z})\}. \qquad (3.22)
\end{aligned}$$

The $\text{PNS}_{\mathbf{z}}$ upper bound is analogously:

$$\begin{aligned}
\text{PNS}_{\mathbf{z}} &\leqslant \sum_{\mathbf{z}} \text{PNS}_{\text{upper-bound}}(\mathbf{z}) \cdot P(\mathbf{z}) \\
&= \sum_{\mathbf{z}} \min\{P(y|x,\mathbf{z}), P(y'|x',\mathbf{z})\} \cdot P(\mathbf{z}) \qquad (3.23) \\
&= \sum_{\mathbf{z}} \min\{P(y|x,\mathbf{z}) \cdot P(\mathbf{z}), P(y'|x',\mathbf{z}) \cdot P(\mathbf{z})\}. \qquad (3.24)
\end{aligned}$$

In comparison with Tian-Pearl bounds:

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

Lemma 3.2.1 makes the superiority of the new bounds clear. The lower bound will be higher than the Tian-Pearl PNS 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 PNS 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.

3.2.3 Graphical Criterion

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

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

In section 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 3.1g and 3.1h. 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 $\text{PNS}_{\mathbf{z}}$ to the Tian-Pearl bounds on PN and PNS.

3.2.4 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 3.1b. 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. These proportions are reflected in the conditional probabilities of table 3.1.

	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 3.1: Conditional probabilities for pain 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,

$$\max\left\{0, \frac{0.5-0.5}{0.5}\right\} \leqslant \mathrm{PN} \leqslant \frac{\min\{0.5, 0.5\}}{0.5},$$

$$0 \leqslant \mathrm{PN} \leqslant 1.$$

Thus, Tian-Pearl bounds provide no information for PN. Utilizing measurements of Z, on the other hand, gives the lower bound from (3.9):

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

The upper bound from (3.12) 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 $\text{PN}_{\mathbf{z}}$ 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 3.5.

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 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 3.2.

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:

$$\max\left\{0, \frac{P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z})}{P(x,y|\mathbf{z})}\right\} \leqslant \text{PN}(\mathbf{z}) \leqslant \min\left\{1, \frac{P(y'_{x'}|\mathbf{z}) - P(x',y'|\mathbf{z})}{P(x,y|\mathbf{z})}\right\}. \tag{3.25}$$

Applying the same technique of summing the minimums and maximums of section 3.2,

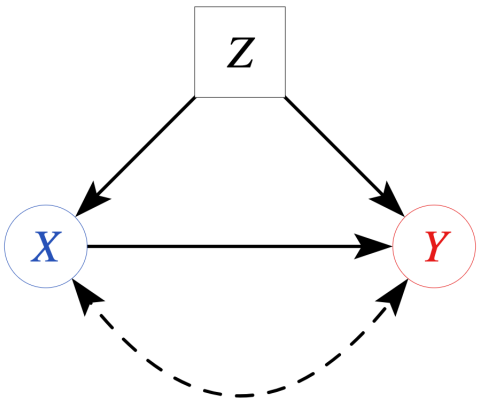


Figure 3.2: Remaining confounding after conditioning on Z.

$\text{PN}_{\mathbf{z}}$ 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 3.2.1, 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 $\forall i : P(x, y|\mathbf{z}) \geqslant P(y'_{x'}|\mathbf{z}) - P(x', y'|\mathbf{z})$.

3.3.2 PNS Bounds

The bounds of PNS($\mathbf{z}$) are:

$$\text{PNS}(\mathbf{z}) \geqslant \max \left\{ \begin{array}{l} 0, \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y|\mathbf{z}) - P(y_{x'}|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y|\mathbf{z}) \end{array} \right\}, \tag{3.26}$$

$$\text{PNS}(\mathbf{z}) \leqslant \min \left\{ \begin{array}{l} P(y_x|\mathbf{z}), \\ P(y'_{x'}|\mathbf{z}), \\ P(x, y|\mathbf{z}) + P(x', y'|\mathbf{z}), \\ P(y_x|\mathbf{z}) - P(y_{x'}|\mathbf{z}) + P(x', y|\mathbf{z}) + P(x, y'|\mathbf{z}) \end{array} \right\}. \tag{3.27}$$

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

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

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

As in the Tian-Pearl PNS bounds of (2.6), the $\text{PNS}_{\mathbf{z}}$ 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 3.3.1. *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:*

$$\max\left\{\sum_i x_{1,i}, \sum_i x_{2,i}, \ldots, \sum_i x_{m,i}\right\} \leqslant \sum_i \max\{x_{1,i}, x_{2,i}, \ldots, x_{m,i}\}. \tag{3.28}$$

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:

$$\min\left\{\sum_i x_{1,i}, \sum_i x_{2,i}, \ldots, \sum_i x_{m,i}\right\} \geqslant \sum_i \min\{x_{1,i}, x_{2,i}, \ldots, x_{m,i}\}. \tag{3.29}$$

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

3.3.3 Graphical Criterion

The graphical criterion for $\mathbf{Z}$ to be advantageous in bounds calculations remains as in section 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

3.3.4 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?

The PNS tells us the answer. 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 3.1b and 3.1d are examples of this scenario. RCT and observational data reveal $P(z) = P(z') = 0.5$ and the conditional probabilities of table 3.2.

	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 3.2: Conditional probabilities for pandemic example

The Tian-Pearl bounds are:

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

$$0.375 \leqslant \text{PNS} \leqslant 0.625.$$

The $\text{PNS}_{\mathbf{z}}$ bounds are:

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

$$= 0.44,$$

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

$$= 0.62.$$

This example demonstrates Tian-Pearl PNS bounds of $0.375 \leqslant \text{PNS} \leqslant 0.625$ and $\text{PNS}_{\mathbf{z}}$ bounds of $0.44 \leqslant \text{PNS} \leqslant 0.62$. 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.

3.4 Violating Additional Information Heuristic

Let us revisit the pandemic example of section 3.3.4 to mirror the discussion in [MLP21]. This time there is only RCT data, referenced in conditional probabilities table 3.3.

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

Table 3.3: Conditional probabilities for pandemic example RCT

Four different bounds can be calculated for PNS:

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

As expected, bounds on PNS_z are narrower than the Tian-Pearl PNS bounds. Surprisingly, if a person knows their ancestry, then their PNS bounds are completely outside the PNS_z bounds. Basically, knowing your ancestry gives you very different, not necessarily narrower, PNS bounds than not knowing your ancestry.

The additional information of ancestral knowledge seems to violate 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 isn't 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 3.4.

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

Table 3.4: Conditional probabilities for coin toss example

Tian-Pearl PNS bounds are $0 \leqslant PNS \leqslant 0.5$ and PNS_z bounds are $0.5 \leqslant PNS \leqslant 0.5$ or $\text{PNS}_z = 0.5$.

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 PNS 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 PNS 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, PNS(heads) = 1 and PNS(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 regardless of coin toss result?" The second question was, "what is the probability of benefiting for a coin toss of heads?"

3.5 Practical Usage

Knowledge of a causal structure enables narrower PN, PS, and PNS 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 3.2.3 and 3.3.3 are satisfied. These are weighted averages of the **Z**-specific probabilities of causation. If an individual's **Z** values are known, the bounds of PN($\mathbf{z}$) and PNS($\mathbf{z}$) in equations 3.7, 3.20, 3.25, 3.26, and 3.27 should be consulted.

Examples in sections 3.2.4 and 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 PNS estimate, so a weighted average of PNS($\mathbf{z}$) will be most accurate and informative. This is analogous to using ACE instead of CACE because of too little data in the conditioning set.

CHAPTER 4

Leveraging Mediation Data

Chapter 3 discussed evaluation of PN, PS, and PNS 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 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 PNS query is posed before treatment is taken, when post-treatment data availability would be rare or impossible.

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. As with the response-type nomenclature of table 2.1, let us repurpose exclusion restriction here. Let M be referred to as a pure mediator when X affects Y only through M. In section 4.2, this assumption will be relaxed. Figure 4.1 depicts example causal graphs of this scenario.

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

Figure 4.1: Mediators where X affects Y only through M.

4.1.1 PNS

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

For simplicity, M is a single binary mediator with values m and m'. Section 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, PNS $= P(y_x, y'_{x'})$, 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 PNS for X and M, $P(m_x, m'_{x'})$. Similarly, the probability of M benefiting Y is the PNS for M and Y, $P(y_m, y'_{m'})$. Another perspective is to use the IV nomenclature of table 2.1. The quantity of interest, PNS, is the probability of a unit or individual being a complier from X to Y. This happens when units or individuals who are compliers from X to M and also compliers from M to Y. Let us call these individuals double-compliers.

The second way X can benefit Y is when X *harms* M and M harms Y. In other words, a unit or individual is a defier from X to M and also a defier from M to Y. Let us call these individuals double-defiers.

Let PN_m and PNS_m be the PN and PNS, respectively, when evaluating with the mediator M. The probability of an individual being a double-complier or a double-defier is the PNS:

$$\begin{aligned}\text{PNS}_m &= P(y_x, y'_{x'}) \\ &= P(y_m, y'_{m'}, m_x, m'_{x'}) + P(y'_m, y_{m'}, m'_x, m_{x'}).\end{aligned} \tag{4.1}$$

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 (4.1):

$$\text{PNS}_m = P(y_m, y'_{m'}) \cdot P(m_x, m'_{x'}) + P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}). \tag{4.2}$$

The Fréchet inequalities for two events will be used in the bound-narrowing techniques below:

$$\max\{0, P(a_1) - [1 - P(a_2)]\} \leqslant P(a_1, a_2) \leqslant \min\{P(a_1), P(a_2)\}. \tag{4.3}$$

Using the following technique, the PNS_m upper bound can sometimes be smaller than the Tian-Pearl upper bound. However, the PNS_m lower bound receives no such advantage.

To see why, let us apply the left side of (4.3) to the probability of being a double-complier:

$$\begin{aligned} P(\text{double-complier}) &= P(y_m, y'_{m'}, m_x, m'_{x'}) \\ &= P(y_m, y'_{m'}) \cdot P(m_x, m'_{x'}) \\ &\geqslant \max\{0, P(y_m) - P(y_{m'})\} \cdot \max\{0, P(m_x) - P(m_{x'})\} \\ &= \max\{0, [P(y_m) - P(y_{m'})] \cdot [P(m_x) - P(m_{x'})]\}. \end{aligned} \tag{4.4}$$

Next, let us apply the left side of (4.3) to the probability of being a double-defier:

$$\begin{aligned} P(\text{double-defier}) &= P(y'_m, y_{m'}, m'_x, m_{x'}) \\ &= P(y'_m, y_{m'}) \cdot P(m'_x, m_{x'}) \\ &\geqslant \max\{0, P(y'_m) - P(y'_{m'})\} \cdot \max\{0, P(m'_x) - P(m'_{x'})\} \\ &= \max\{0, P(y_{m'}) - P(y_m)\} \cdot \max\{0, P(m_{x'}) - P(m_x)\} \\ &= \max\{0, [P(y_{m'}) - P(y_m)] \cdot [P(m_{x'}) - P(m_x)]\} \\ &= \max\{0, [P(y_m) - P(y_{m'})] \cdot [P(m_x) - P(m_{x'})]\}. \end{aligned} \tag{4.5}$$

Equations (4.4) and (4.5) are the same. Lemma 3.2.1 tells us that we can expect no advantage for PNS_m over the Tian-Pearl PNS in this case.

The upper bound, on the other hand, can be lowered. As before, let us start with the

right side of (4.3):

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

Combining $P(\text{double-complier}) + P(\text{double-defier})$ yields the upper bound of PNS_m:

$$\text{PNS}_m \leqslant \min \left\{ \begin{array}{l} P(y_m) \cdot P(m_x) + P(y_{m'}) \cdot P(m_{x'}), \\ P(y_m) \cdot P(m'_{x'}) + P(y_{m'}) \cdot P(m'_x), \\ P(y'_{m'}) \cdot P(m_x) + P(y'_m) \cdot P(m_{x'}), \\ P(y'_{m'}) \cdot P(m'_{x'}) + P(y'_m) \cdot P(m'_x) \end{array} \right\}. \tag{4.6}$$

If there is no pairwise confounding between X, M, and Y, as in figure 4.1a, then this simplifies to observational probabilities:

$$\text{PNS}_m \leqslant \min \left\{ \begin{array}{l} P(y|m) \cdot P(m|x) + P(y|m') \cdot P(m|x'), \\ P(y|m) \cdot P(m'|x') + P(y|m') \cdot P(m'|x), \\ P(y'|m') \cdot P(m|x) + P(y'|m) \cdot P(m|x'), \\ P(y'|m') \cdot P(m'|x') + P(y'|m) \cdot P(m'|x) \end{array} \right\}. \tag{4.7}$$

This upper bound for PNS_m is sometimes worse than Tian-Pearl's PNS upper bound. So, the overall upper bound is:

$$\text{PNS}_m \leqslant \min \left\{ \begin{array}{l} P(y_x), \\ P(y'_{x'}), \\ P(x, y) + P(x', y'), \\ P(y_x) - P(y_{x'}) + P(x', y) \ + P(x, y'), \\ P(y_m) \cdot P(m_x) + P(y_{m'}) \cdot P(m_{x'}), \\ P(y_m) \cdot P(m'_{x'}) + P(y_{m'}) \cdot P(m'_x), \\ P(y'_{m'}) \cdot P(m_x) + P(y'_m) \cdot P(m_{x'}), \\ P(y'_{m'}) \cdot P(m'_{x'}) + P(y'_m) \cdot P(m'_x) \end{array} \right\}. \tag{4.8}$$

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

4.1.2 PN

Under strong exogeneity [TP00], PNS and PN are related with:

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

This means the bounds for PN_m with the models depicted in figures 4.1a and 4.2 are simply the bounds for PNS_m divided by $P(y|x)$.

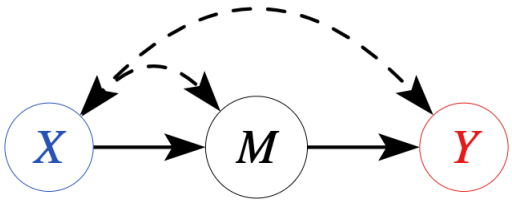


Figure 4.2: Pure mediator with $X \to M$ and $X \to Y$ confounding

However, strong exogeneity does not hold in the graphs of figures 4.1b and 4.2. The reasons will be seen in section 4.1.3. Strong exogeneity does hold in the graph of figure 4.3.

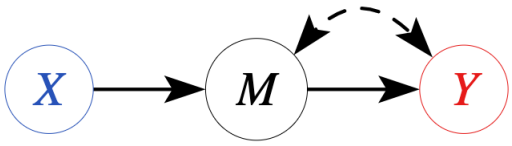


Figure 4.3: Pure mediator with $M \to Y$ confounding

The same bounds were obtained in [DMM17] for PN_m on the simple pure mediator of figure 4.1a.

4.1.3 Graphical Criterion

Two constraints were declared in the derivation to obtain potentially smaller upper bounds on PNS_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) [RR], Twin Networks [BP94], and the Parallel Worlds graph [SP07, 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 4.2 will require multiple hypothetical worlds.

This SWIG is drawn in figure 4.4 for figure 4.1b. 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.

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 4.5 visualizes this counterfactual independency.

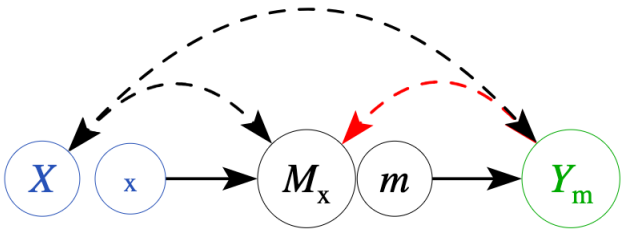


Figure 4.4: Pure mediator SWIG with pairwise confounding

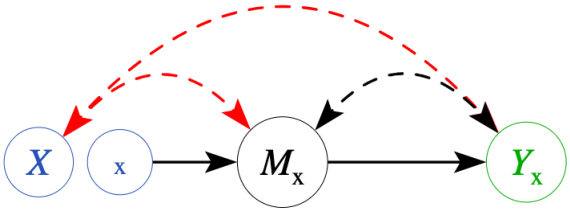


Figure 4.5: Pure mediator SWIG with $Y_x \perp\!\!\!\perp X$ violations in red

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.

4.1.4 Non-binary

The PNS_m upper bound of section 4.1.1 can be generalized to non-binary mediator sets:

$$\begin{aligned}
\text{PNS}_{\mathbf{m}} &= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_{\mathbf{m}_i}, y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \\
&\leqslant \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_{\mathbf{m}_i}), P(y'_{\mathbf{m}_j})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} \qquad (4.9) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}: i<j} \min \left\{ \begin{array}{l} P(y_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{i_x}) + P(y_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_{x'}}), \\ P(y_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{j_{x'}}) + P(y_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{j_x}), \\ P(y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{i_x}) + P(y'_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{i_{x'}}), \\ P(y'_{\mathbf{m}_j}) \cdot P(\mathbf{m}_{j_{x'}}) + P(y'_{\mathbf{m}_i}) \cdot P(\mathbf{m}_{j_x}) \end{array} \right\}. \qquad (4.10)
\end{aligned}$$

Each term in the summation of (4.10) comprises two terms in the summation of (4.9). This is because equation (4.6) 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.

4.1.5 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 4.1a.

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 PNS with PNS_m is straightforward:

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

The PNS_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 4.1a 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.

4.2 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 4.6.

4.2.1 PNS

The following derivation for PNS_m uses three assumptions to obtain measurable probabilities:

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

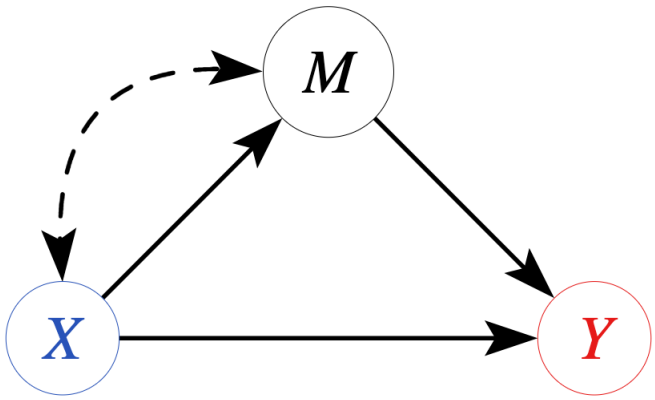


Figure 4.6: Partial mediator M with $X \to M$ confounding

This derivation is equivalent to the proof of theorem 6 in [MLP21]:

$$
\begin{aligned}
\mathrm{PNS}_{\mathbf{m}} &= P(y_x, y'_{x'}) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_x, y'_{x'}, \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) && (4.11) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} P(y_x, y'_{x'} | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \cdot P(\mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}) \\
&\leqslant \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}}), P(y'_{x'} | \mathbf{m}_{i_x}, \mathbf{m}_{j_{x'}})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} && (4.12) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}), P(y'_{x'} | \mathbf{m}_{j_{x'}})\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} && (4.13) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y_x | \mathbf{m}_{i_x}, x), P(y'_{x'} | \mathbf{m}_{j_{x'}}, x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} && (4.14) \\
&= \sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y | \mathbf{m}_i, x), P(y' | \mathbf{m}_j, x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\}.
\end{aligned}
$$

Equation (4.11) is a simple application of total probability, equation (4.12) splits the conditional PNS using the Fréchet upper bound, equation (4.14) 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 (4.14) 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 PNS_m upper bound.

Just like the pure mediator case of section 4.1, sometimes Tian-Pearl upper bounds are smaller than this $\text{PNS}_{\mathbf{m}}$ upper bound. So, the overall upper bound is:

$$\text{PNS}_m \leqslant \min \left\{ \begin{array}{l} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'}) + P(x',y) \; + P(x,y'), \\ \displaystyle\sum_{\mathbf{m}_i, \mathbf{m}_j \in \mathbf{M}} \min\{P(y|\mathbf{m}_i, x), P(y'|\mathbf{m}_j, x')\} \cdot \min\{P(\mathbf{m}_{i_x}), P(\mathbf{m}_{j_{x'}})\} \end{array} \right\}, \quad (4.15)$$

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

4.2.2 PN

As in section 4.1.2, strong exogeneity allows easily computing PN_m from PNS_m by dividing by $P(y|x)$. The $X \to M$ confounding of figure 4.6 does not satisfy strong exogeneity, while the graph in figure 4.7 does.

4.2.3 Graphical Criterion

The following criterion was stated for the derivation of PNS_m in this section:

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

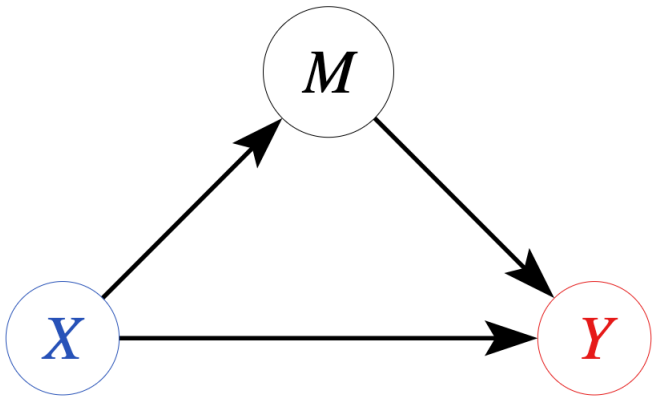


Figure 4.7: Partial mediator M with no confounding among any variable pair

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

The SWIGs in the figures of section 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 4.6 is drawn as a Parallel Worlds graph in figure 4.8.

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.

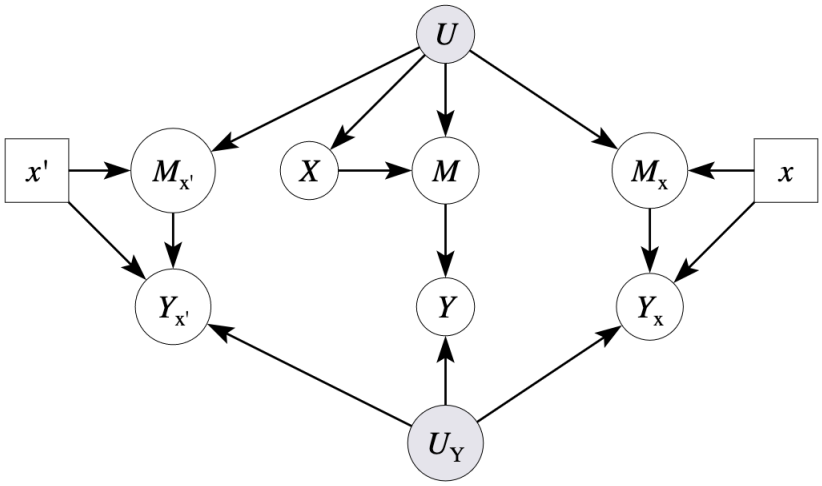


Figure 4.8: Partial mediator Parallel Worlds graph

CHAPTER 5

Leveraging Combinations of Covariates

Chapters 3 and 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 PNS. This chapter will briefly analyze how to overcome criterion violations with mediators and combining covariates and mediators.

5.1 Mediator with Confounding

Figure 5.1 shows a causal graph where the criterion for bounds on $\text{PNS}_{\mathbf{m}}$ using mediator M is not satisfied. In particular, $(Y_m, Y_{m'}) \perp\!\!\!\perp (M_x, M_{x'})$ is violated by the confounder Z.

5.1.1 Pure Mediator

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

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

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

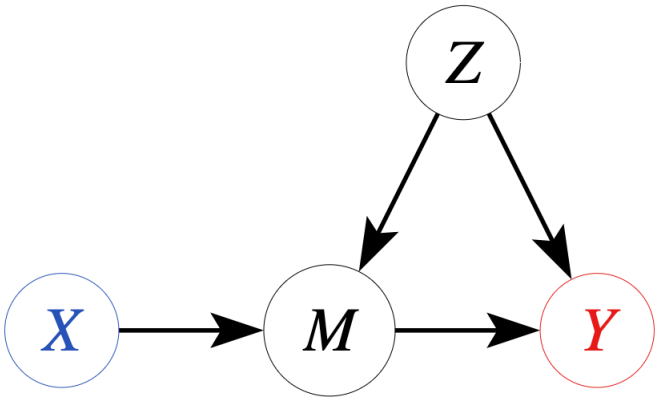


Figure 5.1: Pure mediator with $M \to Y$ confounded by Z

5.1.2 Partial Mediator

Similarly, figure 5.2 shows a partial mediator where the $\text{PNS}_{\mathbf{m}}$ upper bound cannot easily be computed due to the violations of $Y_x \perp\!\!\!\perp M_{x'} \mid M_x$ and $Y_{x'} \perp\!\!\!\perp M_x \mid M_{x'}$.

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

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

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

5.2 Covariates and Mediators

Figure 5.3 presents a causal graph with the possibility of deriving bounds on $\text{PNS}_{\mathbf{m}}$ using the mediator M or the covariate Z.

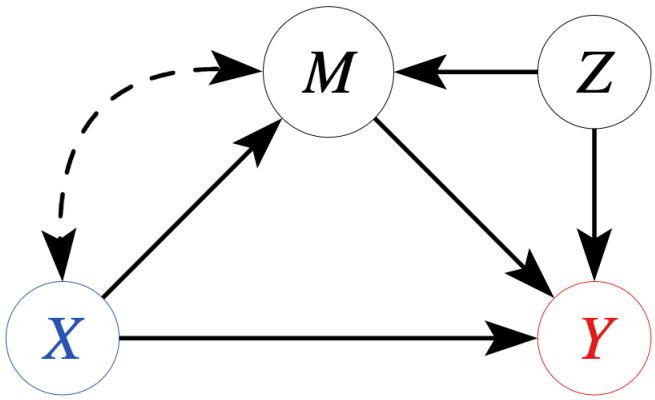


Figure 5.2: Pure mediator with $M \to Y$ confounded by Z

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

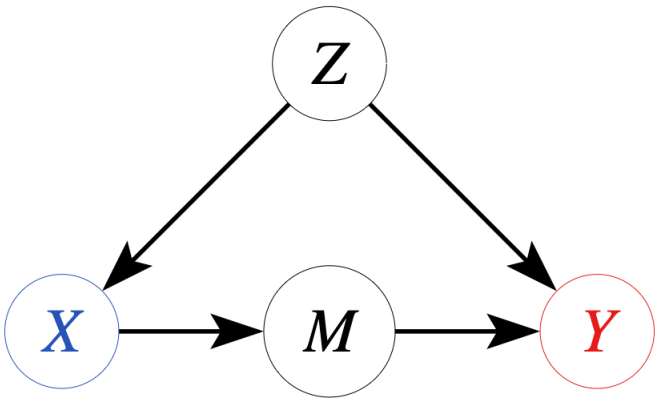


Figure 5.3: Pure mediator M with $X \to Y$ confounded by Z

CHAPTER 6

Conclusion

This thesis analyzed and presented methods of computing narrower bounds on probabilities of causation, PN, PS, and PNS. It demonstrates how significant narrowing of bounds 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 chapter 1, is the difficulty or rarity of obtaining point estimates or sufficiently narrow bounds on CoE. Section 2.7 reviewed how PN, PS, and PNS 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 PNS 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.

Future directions include further refining these techniques, publishing software to compute probability of causation bounds, and integrating with other CoE algorithms and techniques. There are potential gains to be found in utilizing combinations of covariates and mediators beyond what was explored in chapter 5. Software libraries to facilitate these calculations in R and Python would bring CoE analysis to far more researchers and practitioners. Visualizations, like those at `https://learn.ci/pns.html`, allow users to gain an intuition around bounds on probabilities of causation. Finally, valuable CoE work, like Li and Pearl's *Unit Selection based on Counterfactual Logic* [LP19], that depend on PNS bounds would benefit from better accuracy.

REFERENCES

[BP94] Alexander Balke and Judea Pearl. "Probabilistic Evaluation of Counterfactual Queries." *Proceedings of the Twelfth National Conference on Artificial Intelligence*, **1**:230–237, 1994.

[BP13] Alexander Balke and Judea Pearl. "Counterfactuals and policy analysis in structural models." *arXiv preprint arXiv:1302.4929*, 2013.

[CFP20] Carlos Cinelli, Andrew Forney, and Judea Pearl. "A Crash Course in Good and Bad Controls.", 2020. `https://ssrn.com/abstract=3689437`.

[Che97] Patricia W Cheng. "From covariation to causation: A causal power theory." *Psychological review*, **104**(2):367, 1997.

[DMM17] Philip Dawid, Monica Musio, and Rossella Murtas. "The Probability of Causation." *Law, Probability and Risk*, **16**(4):163–179, 2017.

[Gly13] Clark Glymour. "Psychological and normative theories of causal power and the probabilities of causes." *arXiv preprint arXiv:1301.7377*, 2013.

[GP98] David Galles and Judea Pearl. "An axiomatic characterization of causal counterfactuals." *Foundations of Science*, **3**(1):151–182, 1998.

[GP07] Sander Greenland and Judea Pearl. "Causal Diagrams." *Encyclopedia of Epidemiology*, pp. 149–156, 2007.

[Hal00] Joseph Y Halpern. "Axiomatizing Causal Reasoning." *Journal of Artificial Intelligence Research*, **12**:317–337, 2000.

[HR20] MA Hernán and JM Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

[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.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[KFG89] Muin J Khoury, W Dana Flanders, Sander Greenland, and Myron J Adams. "On the measurement of susceptibility in epidemiologic studies." *American Journal of Epidemiology*, **129**(1):183–190, 1989.

[LP19] Ang Li and Judea Pearl. "Unit selection based on counterfactual logic." In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1793–1799. AAAI Press, 2019.

[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.

[MLP21] Scott Mueller, Ang Li, and Judea Pearl. "Causes of Effects: Learning individual responses from population data." *arXiv preprint arXiv:2104.13730*, 2021.

[MP20] Scott Mueller and Judea Pearl. "Which Patients are in Greater Need: A counterfactual analysis with reflections on COVID-19.", Apr 2020. `https://ucla.in/39Ey8sU+`.

[Pea93] J 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, 1995.

[Pea99] Judea Pearl. "Probabilities of Causation: Three counterfactual interpretations and their identification." *Synthese*, **121**(1-2):93–149, 1999.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, Second edition, 2009.

[Pea15] Judea Pearl. "Causes of Effects and Effects of Causes." *Journal of Sociological Methods and Research*, **44**(1):149–164, 2015.

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

[PM21] Judea Pearl and Scott Mueller. "Personalized Decision Making.", Apr 2021. `https://ucla.in/3xFB7ih`.

[PP10] Judea Pearl and Azaria Paz. "Confounding Equivalence in Causal Inference." *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 433–441, 2010.

[RR] Thomas S. Richardson and James M. Robins. "Single World Intervention Graphs: A Primer." `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.644.1881`.

[SGS00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

[SP07] Ilya Shpitser and Judea Pearl. "What Counterfactuals Can Be Tested." *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, 2007.

[SP08] Ilya Shpitser and Judea Pearl. "Complete Identification Methods for the Causal Hierarchy." *Journal of Machine Learning Research*, **9**:1941–1979, 2008.

[TP00] Jin Tian and Judea Pearl. "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence*, **28**(1-4):287–313, 2000.