Learning Causes of Effects and Individual Responses from Population Data and Bayesian Networks*

Scott Mueller, Ang Li, Judea Pearl
{scott, judea}@cs.ucla.edu
ang@cs.fsu.edu

January 28, 2024

Abstract

Identifying causes of observed events is essential in almost every field of science, especially for accurate decision making and generating explanations. The same is true about assessing how an individual would respond to a set of pending interventions. However, such tasks invoke counterfactual relationships and are therefore indeterminable from population data. Even in a fully specified causal Bayesian network, point estimates are generally not estimable for causes of effects or for individual response. For example, the probability of *benefiting* from a treatment concerns an individual having a favorable outcome if treated and an unfavorable outcome if untreated; it cannot be estimated from experimental data, even when conditioned on fine-grained features, because we cannot test both possibilities for the same individual. Tian and Pearl provided bounds on this and other probabilities of causation using a combination of experimental and observational data, yet making no assumption about the structure generating those data. Remarkably, those bounds can be narrowed significantly when structural information is available in the form of a causal model. This paper derives, analyzes, and characterizes these new bounds and illustrates their practical applications in explainable AI, legal responsibility, and personalized medicine.

1 Introduction

Machine learning advances have enabled tremendous capabilities of learning functions accurately and efficiently from enormous quantities of data. These functions allow for better policies, such as whether surgery, chemotherapy, or radiation therapy is most effective for a population of given characteristics such

*An earlier version of this paper was presented in AAAI-23.

as age, sex, and type of symptoms. However, this mapping from characteristics to efficacy can be quite misleading when applied to individual decision making, even when the data originate from a randomized controlled trial (RCT). To see why, let us follow the example treated in [Mueller and Pearl, 2020]. Imagine a novel vaccine for a deadly virus in the midst of a pandemic is in short supply. We want to administer the vaccine to people most likely to benefit from it. In other words, we need to identify the group most likely to *both* survive if vaccinated and succumb if unvaccinated.

A clinical study is conducted to test the effectiveness of the vaccine. For simplicity, let's assume a binary age classification: young (sixty years old and under) and old (over sixty years old). Older people survive 57% of the time when vaccinated and 37% of the time when unvaccinated, while younger people survive 55% of the time when vaccinated and 45% of the time when unvaccinated. A naïve interpretation is that the vaccine is 10 percentage points more effective for older people and, therefore, they should be vaccinated first.

However, a different picture emerges if we assess the percentage of beneficiaries in the two groups. These percentages, known as Probability of Necessity and Sufficiency (PNS) [Pearl, 1999], can be tightly bound [Tian and Pearl, 2000] and falls, given the data above, between 20% and 57% for the older patients and between 10% and 55% for the younger patients. We see that it's anything but clear which group should be vaccinated first.

What is more remarkable is these bounds can be narrowed significantly if data from observational studies is also available, and may even flip priority from the elderly to the young. Observational studies reflect outcomes for individuals who decide on their own whether to get vaccinated or not. In our example, one can show that the bounds for over-sixties and under-sixties may become $[20\%, 40\%]$ and $[40\%, 55\%]$, respectively, thus reversing the naïve priorities above, and clearly show priority to vaccinate the young, not the elderly.

Since Tian and Pearl [Tian and Pearl, 2000], the problem of bounding probabilities of causation was analyzed by combining only two sources of information: experimental data and observational studies, making no assumptions whatsoever about the model generating the data. This paper shows[1] that, surprisingly, knowing the structure of the causal graph allows us to narrow these bounds, despite the fact that the graph may seem redundant; i.e., we already know the causal effects. Moreover, the graph adds information about an individual, although it describes properties of the population. A fully specified causal Bayesian network allows experimental results to be derived from observational data, but historically has lacked the ability to produce narrower bounds on probabilities of causation than Tian and Pearl's bounds. Knowledge of the causal structure and data allows us to narrow these bounds because we can then partition bounds on subsets of covariates and mediators, obtain local bounds on the partitions, and combine the bounds. This partitioning gives us a finer-grained perspective on possible values for probabilities of causation. The analysis of causes of effects can now take advantage of the causal diagram.

[1] Supplementary material is available at https://ftp.cs.ucla.edu/pub/stat_ser/r505-sup.pdf

2 Preliminaries and Related Work

In this section, we review the definitions for the three aspects of causation as defined in [Pearl, 1999]. We use the causal diagrams [Koller and Friedman, 2009; Pearl, 1995, 2009; Spirtes et al., 2000] and the language of counterfactuals in its structural model semantics, as given in [Balke and Pearl, 2013; Galles and Pearl, 1998; Halpern, 2000].

We use $Y_x = y$ to denote the counterfactual sentence "Variable Y would have the value y, had X been x." For simplicity purposes, in the rest of the paper, we use y_x to denote the event $Y_x = y$, $y_{x'}$ to denote the event $Y_{x'} = y$, y'_x to denote the event $Y_x = y'$, and $y'_{x'}$ to denote the event $Y_{x'} = y'$. For notational simplicity, we limit the discussion to binary X and Y.

Three prominent probabilities of causation are the following:

Definition 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 [Pearl, 1999]*

$$\begin{aligned} PN &\triangleq P(Y_{x'} = false | X = true, Y = true) \\ &\triangleq P(y'_{x'} | x, y) \end{aligned} \tag{1}$$

Definition 2 (Probability of sufficiency (PS)). *[Pearl, 1999]*

$$PS \triangleq P(y_x | y', x') \tag{2}$$

Definition 3 (Probability of necessity and sufficiency (PNS)). *[Pearl, 1999]*

$$PNS \triangleq P(y_x, y'_{x'}) \tag{3}$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

Tian and Pearl [Tian and Pearl, 2000] provide tight bounds for PNS, PN, and PS without a causal diagram using Balke's program [Balke and Pearl, 1997] (we will call them Tian-Pearl bounds). Li and Pearl [Li and Pearl, 2019] provide a theoretical proof of the tight bounds for PNS, PS, PN, and other probabilities of causation without a causal diagram.

PNS, PN, and PS have the following tight bounds:

$$\text{PNS} \geq \max \left\{ \begin{array}{c} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{array} \right\} \tag{4}$$

$$\text{PNS} \leq \min \left\{ \begin{array}{c} P(y_x) \\ P(y'_{x'}) \\ P(x,y) + P(x',y') \\ P(y_x) - P(y_{x'}) + \\ + P(x,y') + P(x',y) \end{array} \right\} \tag{5}$$

$$\text{PN} \geq \max \left\{ \begin{array}{c} 0 \\ \frac{P(y) - P(y_{x'})}{P(x,y)} \end{array} \right\} \tag{6}$$

$$\text{PN} \leq \min \left\{ \begin{array}{c} 1 \\ \frac{P(y'_{x'}) - P(x',y')}{P(x,y)} \end{array} \right\} \tag{7}$$

Note that we only consider PNS and PN here because the bounds of PS can be easily obtained by exchanging x with x' and y with y' in the bounds of PN.

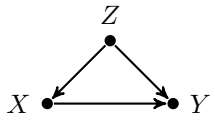
To obtain bounds for a specific population, defined by a set C of characteristics, the expressions above should be modified by conditioning each term on $C = c$. In this paper, however, we obtain narrower bounds of PNS by leveraging another source of knowledge – the causal diagram behind the data, together with measurements of a set Z of covariates in that diagram. We provide graphical conditions under which the availability of such measurements would improve the bounds and demonstrate, both analytically and by simulation, the degree of improvement achieved. Narrower bounds and graphical criteria can be obtained for PN and PS through the same mechanism detailed in the proofs in the appendix.

Experimental probabilities, such as $P(y_x)$, can be easily computed from observational data in fully specified causal Bayesian networks. This is accomplished through a graph mutilation [Pearl, 2009, §1.3.1] by removing inbound arrows to X and setting $X = x$.

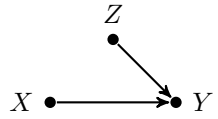
3 Bounds with Causal Diagram

3.1 Non-descendant Covariates

Theorems 4 and 5 below provide bounds for PNS when a set $\mathbf{Z}$ of variables can be measured which satisfy only one simple condition: $\mathbf{Z}$ contains no descendants of X. This condition is important because if X was set to x and $\mathbf{Z}$ contains a descendant of X, then $\mathbf{Z}$ could be altered as well and $P(y_x|\mathbf{z})$ would be unmeasurable. This unmeasurability is clear in a causal Bayesian network. After conditioning on $\mathbf{z}$, Y is now dependent on both X being set to x and Z_x being set to z'. These types of counterfactual probabilities are not estimable with causal Bayesian networks. If the descendant is independent of Y_x, then $P(y_x|z)$ would be measurable, but that descendant wouldn't contribute to any narrowing of bounds. These bounds are always contained within the Tian-Pearl bounds of equations 4, 5, 6, and 7.



(a) Confounder Z



(b) Outcome-affecting covariate Z

Figure 1: Z is not a descendant of X

Theorem 4. *Given a causal diagram G and distribution compatible with G, let Z be a set of variables that does not contain any descendant of X in G, then PNS is bounded as follows:*

$$PNS \geq \sum_z \max \left\{ \begin{array}{c} 0, \\ P(y_x|z) - P(y_{x'}|z), \\ P(y|z) - P(y_{x'}|z), \\ P(y_x|z) - P(y|z) \end{array} \right\} \times P(z) \tag{8}$$

$$PNS \leq \sum_z \min \left\{ \begin{array}{c} P(y_x|z), \\ P(y'_{x'}|z), \\ P(x,y|z) + P(x',y'|z), \\ P(y_x|z) - P(y_{x'}|z) \\ +\ P(x,y'|z) + P(x',y|z) \end{array} \right\} \times P(z) \tag{9}$$

Proof. See Appendix. □

Note that, unlike the subpopulation bounds, where each term is conditioned on $C = c$, here PNS is *not* conditioned on $Z = z$. This is because the measurement of Z is conducted in the study, but may not be available for the individual seeking advice. Examples are illustrated in Section 4.

Note also that if only experimental data are available (i.e., $P(Y), P(Y,X), P(Y|Z), P(Y,X|Z)$ are not measured), arguments to the max or min functions involving observational data can be disregarded. For example, the lower bounds of Theorem 4 would become $\max\{P(y_x) - P(y_{x'}), \sum_z \max\{0, P(y_x|z) - P(y_{x'}|z)\} \times P(z)\}$.

3.1.1 Sufficient Covariates

In Figures 1a and 1b, Z is not a descendant of X and additionally satisfies the back-door criterion. For such cases the PNS bounds can be simplified:

Theorem 5. *Given a causal diagram G and distribution compatible with G, let Z be a set of variables satisfying the back-door criterion [Pearl, 1993] in G, then the PNS is bounded as follows:*

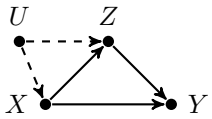


Figure 2: Mediator Z with direct effect

$$PNS \geq \sum_z \max\{0, P(y|x,z) - P(y|x',z)\} \times P(z) \tag{10}$$

$$PNS \leq \sum_z \min\{P(y|x,z), P(y'|x',z)\} \times P(z) \tag{11}$$

Proof. See Appendix. □

The significance of Theorem 5 is due to the ability to compute bounds using purely observational data.

3.2 Mediation

3.2.1 Partial Mediator

In Figure 2, Z is a descendant of X, so we cannot use Theorems 4 and 5. However, the absence of confounders between Z and Y and between X and Y permits us to bound PNS as follows:

Theorem 6. *Given a causal diagram G and distribution compatible with G, let Z be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp\!\!\!\perp X \cup Z_{x'} \mid Z_x)$ in G, then the PNS is bounded as follows:*

$$PNS \geq \max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \tag{12}$$

$$PNS \leq \min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ \\ P(y_x) - P(y_{x'}) \\ + \ P(x,y') + P(x',y), \\ \\ \sum_z \sum_{z'} \min\{P(y|x,z), \\ P(y'|x',z')\} \times \\ \min\{P(z_x), P(z'_{x'})\} \end{array} \right\} \tag{13}$$

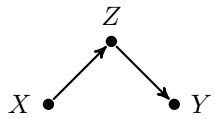


Figure 3: Mediator Z with no direct effect

Proof. See Appendix. □

Note that although this lower bound is unchanged from Tian and Pearl, the upper bound contains a vital additional argument to the min function. This new term can significantly reduce the upper bound. The rest of the terms are included because sometimes Tian and Pearl's bounds are superior. The following Theorem has the same quality.

3.2.2 Pure Mediator

Figure 3 is a special case of Figure 2, in which X has no direct effect on Y. The resulting bounds for PNS read:

Theorem 7. *Given a causal diagram G in Figure 3 and distribution that compatible with G, then PNS are bounded as follow:*

$$PNS \geq \max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \tag{14}$$

$$PNS \leq \min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ \\ P(y_x) - P(y_{x'}) \\ + \ P(x',y) + P(x,y'), \\ \\ \Sigma_z \Sigma_{z' \neq z} \min\{P(y|z), P(y'|z')\} \times \\ \min\{P(z|x), P(z'|x')\} \end{array} \right\} \tag{15}$$

Proof. See Appendix. □

The core terms for Theorems 6 and 7 added to the upper bounds notably only require observational data.

	Drug	No Drug
Women	1 out of 110 recovered (1%)	13 out of 120 recovered (11%)
Men	313 out of 354 recovered (88%)	114 out of 116 recovered (98%)
Overall	314 out of 464 recovered (68%)	127 out of 236 recovered (54%)

Table 1: Results of a drug study with gender taken into account

4 Examples

4.1 Credit to the Treatment

The manufacturer of a drug wants to claim that a non-trivial number of recovered patients who were given access to the drug owe their recovery to the drug. So they conduct an observational study; they record the recovery rates of 700 patients. 464 patients chose to take the drug and 236 patients did not. The results of the study are in table 1. The manufacturer claims success for their drug because the overall recovery rate from the observational study has increased from 54% to 68% for non-drug-takers to drug-takers.

The number of recovered patients that should credit the drug for their recovery are those who would recover if they had taken the drug and would not recover if they had not taken the drug. This is the PNS.

Let $X = x$ denote the event that the patient took the drug and $X = x'$ denote the event that the patient did not take the drug. Let $Y = y$ denote the event that the patient has recovered and $Y = y'$ denote the event that the patient has not recovered. Let $Z = z$ represent female patients and $Z = z'$ represent male patients. Suppose we know an additional fact, estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug. Additionally, as we can see from the data, men are significantly more likely to take the drug than women are. The causal diagram is shown in Figure 1a.

Node Z on the graph satisfies the back-door criterion, therefore we can compute the causal effect $P(y_x)$ and $P(y'_x)$ via the adjustment formula [Pearl, 1993] and observational data from table 1, where,

$$P(y_x) = \sum_z P(y|x,z)P(z) = 0.597,$$
$$P(y_{x'}) = \sum_z P(y|x',z)P(z) = 0.696,$$
$$P(y'_{x'}) = 1 - P(y_{x'}) = 0.304.$$

Therefore, the bounds of PNS computed using equations 4 and 5 are $0 \leq PNS \leq 0.297$, where the diagram was used only to identify the causal effects y_x

and $y_{x'}$. These bounds aren't informative enough to conclude whether or not the drug was the cause of recovery for a meaningful number of patients. They suggest that the fraction of beneficiaries can be as low as 0% or as high as 29.7%. Now, consider the bounds in Theorem 5 which takes into account the position of Z in the diagram. Since Z satisfies the back-door criterion, we can use equations 10 and 11 to compute $0 \leqslant PNS \leqslant 0.01$. The conclusion now is obvious. At most 7 out of 314 patients' recoveries can be credited to the drug. This is strong evidence that counters the manufacturer's claim.

4.2 Inflammation Mediator

As before, let X and Y represent drug consumption and recovery. Let Z represent acute inflammation with z being present and z' being absent. The drug reduces inflammation. However, in some people the drug causes acute inflammation, which has adverse effects on recovery. The causal structure is depicted in Figure 3. We observe the following proportions among drug takers, non-takers, with inflammation, and without inflammation:

$$\begin{aligned} P(y|z) &= 0.5, & P(z|x) &= 0.1, \\ P(y|z') &= 0.5, & P(z|x') &= 0.1. \end{aligned}$$

The Tian-Pearl PNS upper bound is:

$$\text{PNS} \leqslant \min\{P(y|x), P(y'|x')\} = 0.5.$$

Given that the lower bound is 0, these bounds are not very informative. If we knew that an individual would react to the drug with acute inflammation, we would only look at the data comprising of people reacting to the drug with acute inflammation. Since we are conditioning on z, $PNS = 0$ because the outcome, Y, will have the same result regardless of whether the person consumed the drug. So knowing a person's inflammation response to the drug narrows PNS from a wide $[0, 0.5]$ to a point estimate of 0. Imagine, for this drug, that we can't know ahead of time how a person will react inflammation-wise. We can only observe acute inflammation after the drug is administered. Since we have population data from patients who have already taken the drug, we can utilize this mediator to bound the PNS for new patients who haven't yet taken the drug:

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

The mediator-improved PNS upper bound is significantly smaller than what the Tian-Pearl upper bound provides, 0.1 vs 0.5. The new upper bound can now be effectively weighed against other factors like cost and side-effects.

4.3 Ancestral Covariate

Let's continue from the introduction, where X represents vaccination with x being vaccinated and x' being unvaccinated and Y represents survival with y is surviving and y' is succumbing to the pandemic. Instead of classifying by age, let's assume our machine learning algorithm uncovers a correlation between survival and ancestry. Let Z represent ancestry and, for simplicity, there are only two ancestries, z and z'. Either graph of Figure 1 is representative of this. Our RCT data reveals:

$$\begin{aligned} P(Z=z) &= 0.5, \\ P(y_x|Z=z) &= 0.75, \\ P(y_{x'}|Z=z) &= 0.2, \end{aligned} \qquad \begin{aligned} P(y_x|Z=z') &= 0.25, \\ P(y_{x'}|Z=z') &= 0.6. \end{aligned}$$

We now have four different bounds on PNS:

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

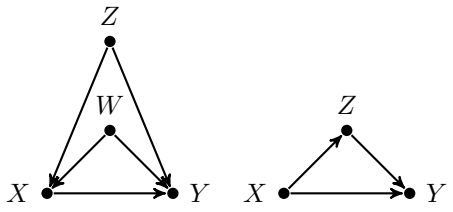
As expected, using the causal diagram and ancestral Z yields narrower bounds than the Tian-Pearl bounds. However, it's surprising that knowing a person has either ancestry z or z' gives us bounds outside of our new bounds. In fact, they are completely outside the wider Tian-Pearl bounds. This is discussed in section 6.

In the meantime, it's important to recognize that the last two ancestry-specific PNS bounds are what should be referred to if an individual knows their ancestry. The covariate-improved PNS bounds should only be referred to if a person's ancestry is unknown. This might be because the person was adopted with no hint as to whether they're from ancestry z or z' (physical features are right in between or indistinguishable).

5 Simulation Results

We randomly generated 100000 sample distributions compatible with each the causal diagrams depicted in Figures 4a, 1a, 4b, and 3 for Theorems 4, 5, 6, and 7, respectively. Given sample distribution i, let $[a_i, b_i]$ be the bounds utilizing the proposed Theorems and $[c_i, d_i]$ be the traditional Tian-Pearl bounds [Li and Pearl, 2021]. The following is computed for each causal diagram as summarized in Table 2:

- Average increased PNS lower bound: $\frac{\sum(a_i - c_i)}{100000}$
- Average decreased PNS upper bound: $\frac{\sum(d_i - b_i)}{100000}$



(a) Z is a non-descendant of X (b) Z is a partial mediator

Figure 4: Causal diagrams for simulation

	Incr'd lower bound	Decr'd upper bound	Tian-Pearl PNS gap	Theorems PNS gap	Samples benefiting
Non-desc	0.0238	0.0237	0.2673	0.2197	85.622%
Suff covar	0.0266	0.0264	0.2197	0.1668	75.025%
Part med	0.0000	0.0047	0.2289	0.2242	12.532%
Part med 2	0.0000	0.0382	0.2768	0.2386	100.00%
Pure med	0.0000	0.0935	0.2605	0.1670	100.00%

Table 2: Performance metrics for Theorems 4 (Non-desc), 5 (Suff covar), 6 (Part med & Part med 2), and 7 (Pure med)

- Average gap in Tian-Pearl PNS bounds: $\frac{\sum(d_i - c_i)}{100000}$
- Average gap utilizing Theorems 4, 5, 6, and 7: $\frac{\sum(b_i - a_i)}{100000}$
- Number of sample distributions benefiting from Theorems 4, 5, 6, and 7: $\sum e_i$, where $e_i = 1$ if $(a_i > c_i)$ or $(b_i < d_i)$, $e_i = 0$ otherwise.

For each causal diagram, 100 out of 100000 sample distributions are randomly selected, sorted by lower and upper PNS bound, and then drawn with and without considering the causal diagram (Figures 5 to 8).

Table 2 shows the average gaps between Tian-Pearl PNS bounds and Theorem 6's bounds are similar for the partial mediator of Figure 4b (Part med in Table 2). This is because only 12.532% of samples are narrowed by the proposed Theorem 6. A second set of sample distributions were generated repeatedly until 100000 narrowed samples were available (Part med 2 in Table 2). This time the difference in gaps were significant, which is important if the costs of including partial mediator data are low.

6 Discussion

We have shown that knowledge of a causal structure enables narrower PNS bounds to be estimated, compared with the tight bounds of Tian and Pearl

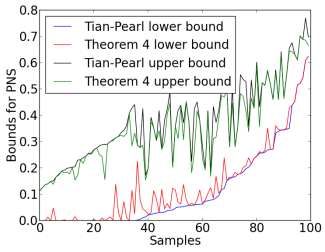


Figure 5: PNS bounds for causal diagram of Figure 4a

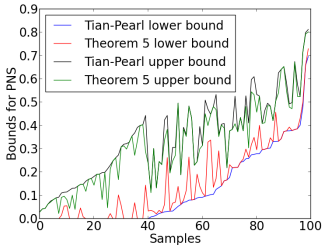


Figure 6: PNS bounds for causal diagram of Figure 1a

which were derived without such knowledge. However, it must be emphasized that this narrowing is only applicable to individuals whose Z characteristics are not known at decision time. If their Z values are known, the bounds of equations 4 and 5, conditioned on those values, should be consulted. Example 4.3 provides a scenario where people who know their ancestry have very different PNS bounds than people who don't know their ancestry. You would expect the additional information of ancestral knowledge would further narrow the bounds, but they change the bounds to a different non-overlapping range. This violates the heuristic that *additional information* should narrow the bounds or, at worst, not widen them. To rephrase, if you don't know someone's ancestry, the probability they benefit from this drug is between 0.275 and 0.5. Once you acquire the additional information that the person is of ancestry z, the probability they benefit from this treatment becomes between 0.55 and 0.75. How is this possible? Was the person's probability of benefiting never really between 0.275 and 0.5 that we calculated before knowing their ancestry?

The reason for this seeming inconsistency is that we're asking different questions. When we didn't know the ancestry, we were asking, "what is the probability of benefiting for a person regardless of ancestry?" When we found out the person is of ancestry z, we then asked a different question, "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 Pearl,

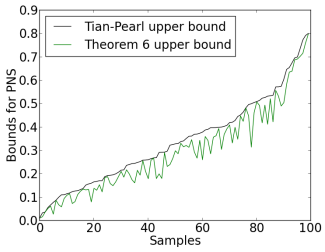


Figure 7: PNS bounds for causal diagram of Figure 4b among narrowed samples referenced by *part med 2* in Table 2

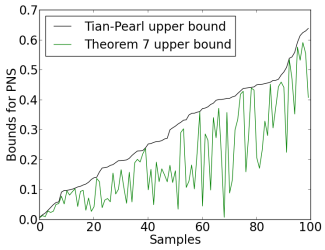


Figure 8: PNS bounds for causal diagram of Figure 3

2009, p. 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's assume the following measurements are taken:

$$\begin{aligned} &P(y_x) = 0.5, && P(y_{x'}) = 0.5, \\ &P(y_x|Z = \text{heads}) = 1, && P(y_{x'}|Z = \text{heads}) = 0, \\ &P(y_x|Z = \text{tails}) = 0, && P(y_{x'}|Z = \text{tails}) = 1. \end{aligned}$$

Tian-Pearl bounds gives us $0 \leqslant PNS \leqslant 0.5$ and the bounds utilizing Z are $0.5 \leqslant PNS \leqslant 0.5$ or PNS $= 0.5$.

Now, let's 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, Z, the odds of winning \$1 are 50/50 whether you bet on heads or tails. PNS is also 0.5 because benefiting from betting on heads is true only when the coin toss was heads. 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?"

7 Conclusion

In this work, we have developed a graphical method of learning individualized functions (representing PNS, PN, and PS) from population data, based on the structure of the causal graph. These methods generalize the PN, PS, and PNS bounds derived in [Tian and Pearl, 2000], the bounds derived in [Kuroki and Cai, 2011], and the PN bounds derived in [Dawid et al., 2017]. Often these functions return bounds rather than point estimates. This paper shows nevertheless that the bounds obtained can be quite informative. Machine learning algorithms can easily incorporate these techniques to achieve both data interpretability and decision making accuracy for situation-specific cases.

Acknowledgments

This research was supported in parts by grants from the National Science Foundation [#IIS-2106908], Office of Naval Research [#N00014-17-S-12091 and #N00014-21-1-2351], and Toyota Research Institute of North America [#PO-000897].

References

Balke, A., & Pearl, J. (1997). *Probabilistic counterfactuals: Semantics, computation, and applications* [Doctoral dissertation, University of California, Los Angeles].

Balke, A., & Pearl, J. (2013). Counterfactuals and policy analysis in structural models. *arXiv preprint arXiv:1302.4929*.

Dawid, P., Musio, M., & Murtas, R. (2017). The probability of causation. *Law, Probability and Risk*, *16*(4), 163–179.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*(1), 151–182.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, *12*, 317–337.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Kuroki, M., & Cai, Z. (2011). Statistical analysis of 'probabilities of causation' using co-variate information. *Scandinavian Journal of Statistics*, *38*(3), 564–577. https://doi.org/https://doi.org/10.1111/j.1467-9469.2011.00730.x

Li, A., & Pearl, J. (2019). Unit selection based on counterfactual logic. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1793–1799.

Li, A., & Pearl, J. (2021). Unit selection with causal diagram. *arXiv preprint arXiv:2109.07556*.

Mueller, S., & Pearl, J. (2020). Which Patients are in Greater Need: A counterfactual analysis with reflections on COVID-19 [Accessed: 2022-06-05].

Pearl, J. (1993). Aspects of graphical models connected with causality. *Proceedings of the 49th Session of the International Statistical Institute, Italy*, 399–401.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

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

Pearl, J. (2009). *Causality* (Second). Cambridge University Press.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1-4), 287–313.