Causes of Effects: Learning Individual Responses from Population Data

Scott Mueller, Ang Li and Judea Pearl
Cognitive Systems Laboratory, Computer Science Department, University of California Los Angeles
{scott, angli, judea}@cs.ucla.edu

Abstract
The problem of individualization is crucial in almost every field of science. Identifying causes of specific observed events is likewise essential for accurate decision making as well as explanation. However, such tasks invoke counterfactual relationships, and are therefore indeterminable from population data. 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 an individual. Tian and Pearl provided bounds on this and other probabilities of causation using a combination of experimental and observational data. Those bounds, though tight, can be narrowed significantly when structural information is available in the form of a causal model. This added information may provide the power to solve central problems, such as explainable AI, legal responsibility, and personalized medicine, all of which demand counterfactual logic. This paper derives, analyzes, and characterizes these new bounds, and illustrates some of their practical applications.

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, like whether a surgery, chemotherapy, or radiation therapy is most effective for a population of given characteristics such 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’s 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: sixty years old and under and 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 bounded [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 shows1 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. Knowledge of the causal structure and data allows us to narrow 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.

1Supplementary material is available at https://ftp.cs.ucla.edu/pub/stat_set/r505-sup.pdf
The analysis of causes of effects can now take advantage of the causal diagram.

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 [Pearl, 1995; Spirtes et al., 2000; Pearl, 2009; Koller and Friedman, 2009] 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 $Y_x = y'$. For notational simplicity, we limit the discussion to binary $X$ and $Y$. Extension to multi-valued variables are straightforward [Pearl, 2009].

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]

$$PN \triangleq P(Y_{x'} = false|X = true, Y = true)$$

$$PN \triangleq P(y'_{x}|x, y)$$

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

$$PS \triangleq P(y|x', x')$$

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

$$PNS \triangleq P(y, y'_x)$$

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:

$$PNS \geq \max \left\{ \frac{0}{P(y) - P(y'|y)} \right\}$$

$$PNS \geq \max \left\{ \frac{P(y_{x'}) - P(y_{x'|y'})}{P(y) - P(y')} \right\}$$

$$PNS \leq \min \left\{ \frac{P(y_{x'})}{P(y')} \right\}$$

$$PNS \leq \min \left\{ \frac{P(x, y) + P(x', y')}{P(y_{x'})} \right\}$$

$$PNS \leq \min \left\{ \frac{P(x, y') + P(x', y)}{P(y_{x'})} \right\}$$

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.

3 Bounds with Causal Diagram

3.1 Non-descendant Covariates

Theorems 4 and 5 below provide bounds for PNS when a set $Z$ of variables can be measured which satisfy only one simple condition: $Z$ contains no descendants of $X$. This condition is important because if $X$ was set to $x$ and $Z$ contains a descendant of $X$, then $Z$ could be altered as well and $P(y_{x}|z)$ would be unmeasurable. 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.

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:

$$PNS \geq \sum_{z} \max \left\{ \begin{array}{l} P(y_{x}|z) - P(y_{x'|z}), \\ P(y|z) - P(y_{x}|z), \\ P(y_{x'}|z) - P(y|z) \end{array} \right\} \times P(z) \quad (8)$$

$$PNS \leq \sum_{z} \min \left\{ \begin{array}{l} P(y_{x}|z), \\ P(y_{x'|z}), \\ P(x, y|z) + P(x', y'|z), \\ P(y_{x'}|z) - P(y_{x}|z) + P(x, y'|z) \end{array} \right\} \times P(z) \quad (9)$$

Proof. See Appendix.

2713
Given a causal diagram $G$, let $Z$ be a set of variables such that $\forall x, x' \in X : x \neq x', (Y_x \perp X \cup Z_{x'} | Z_x)$ in $G$, then the PNS is bounded as follows:

$$PNS \geq \max \left\{ 0, P(y_x) - P(y_{x'}), P(y) - P(y_{x'}), P(y_x) - P(y) \right\}$$

(12)

$$PNS \leq \min \left\{ P(y_x) - P(y_{x'}), P(x, y) + P(x', y'), \sum_z \min\{P(y|x, z), P_i(y|^{x', z'})\} \times \min\{P(z), P(z')\} \right\}$$

(13)

**Proof.** See Appendix. 

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

### 3.2 Mediation

#### 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:

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

(10)

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

(11)

**Proof.** See Appendix. 

The significance of Theorem 5 is due to the ability to compute bounds using purely observational data.
### Table 1: Results of a drug study with gender taken into account

<table>
<thead>
<tr>
<th></th>
<th>Drug</th>
<th>No Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>1 out of 110</td>
<td>13 out of 120</td>
</tr>
<tr>
<td></td>
<td>recovered (1%)</td>
<td>recovered (11%)</td>
</tr>
<tr>
<td>Men</td>
<td>313 out of 354</td>
<td>114 out of 116</td>
</tr>
<tr>
<td></td>
<td>recovered (88%)</td>
<td>recovered (98%)</td>
</tr>
<tr>
<td>Overall</td>
<td>314 out of 464</td>
<td>127 out of 236</td>
</tr>
<tr>
<td></td>
<td>recovered (68%)</td>
<td>recovered (54%)</td>
</tr>
</tbody>
</table>

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

### 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 $P(y|x_z)$ and $P(y'|x'_z)$ via the adjustment formula [Pearl, 1993] and observational data from Table 1, where,

$$P(y_z) = \sum_z P(y|x, z)P(z) = 0.597,$$

$$P(y_{z'}) = \sum_z P(y|x', z)P(z) = 0.696,$$

$$P(y'_{z'}) = 1 - P(y_{z'}) = 0.304.$$

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 $y_z$ and $y_{z'}$. 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 $0 \leq PNS \leq 0.01$. 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:

$$P(y|z) = 0.5, \quad P(z|x) = 0.1,$$

$$P(y|z') = 0.5, \quad P(z|x') = 0.1.$$

The Tian-Pearl PNS upper bound is:

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

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:

$$PNS \leq \min \left\{ P(y|z) \cdot P(z|x) + P(y|z') \cdot P(z|x') ,
            P(y|z) \cdot P(z'|x) + P(y|z') \cdot P(z'|x') ,
            P(y'|z') \cdot P(z|x) + P(y'|z) \cdot P(z|x') ,
            P(y'|z') \cdot P(z'|x') + P(y'|z) \cdot P(z'|x) \right\} = 0.1.$$

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
We randomly generated 100000 samples, as expected, using the causal diagram and ancestral distribution. This might be because the person was adopted with each the causal diagrams depicted in Figures 4a, 1a, 5 Simulation Results

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

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

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

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 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 physical features. As expected, 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: \(\frac{\sum (d_i - b_i)}{100000}\)
- 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.

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).
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 regardless of coin toss result?” The second question was, “what is the probability of benefiting from betting on heads when we didn’t know the coin toss result?” The second question is, “what is the 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 from betting on heads when we knew the coin toss resulted in 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.

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


