Alternative Measures of Direct and Indirect Effects

Jose M. Peña JOSE.M.PENA@LIU.SE
STIMA, IDA, Linköping University, Sweden

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

There are a number of measures of direct and indirect effects in the literature on causality. These are suitable in some cases and unsuitable in others. We describe a case where the existing measures are unsuitable and propose new suitable ones. We also show that the new measures can partially handle unmeasured treatment-outcome confounding, and bound long-term effects by combining experimental and observational data. We also introduce the concepts of indirect benefit and harm (i.e., through a mediator), and use our new measure to quantify them.

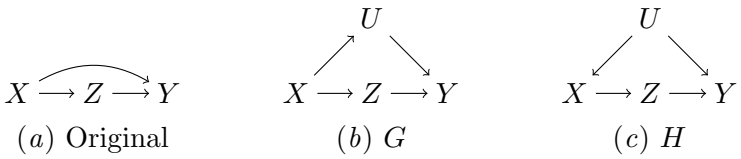
Keywords: Direct effect; indirect effect; total effect; probability of benefit; probability of harm.

1. Introduction

Consider the causal graph in Figure 1(*a*), which is studied by Pearl (2001, 2009) and where X, Z and Y represents an applicant's gender, qualifications and hiring, respectively. Let X be binary taking values in $\{x, x'\}$. Let $Y_{\dot{x}}$ denote the counterfactual value of Y when X is set to value $\dot{x} \in \{x, x'\}$. Likewise for $Z_{\dot{x}}$. Let $Y_{\dot{x},Z_{\ddot{x}}}$ denote the counterfactual value of Y when X is set to value $\dot{x} \in \{x, x'\}$ and Z is set to value $Z_{\ddot{x}}$ with $\ddot{x} \in \{x, x'\}$. The edge $X \to Y$ in the graph in Figure 1(*a*) represents that the hirer questions applicants about their gender, and that their answers may have an effect on hiring them. Pearl imagines a policy maker who may be interested in predicting the gender mix in the work force, if it were illegal for the hirer to question applicants about their gender. This quantity corresponds to the effect of gender on hiring mediated by qualifications. Pearl argues that the answer to this question lies in deactivating the direct path $X \to Y$. He also argues that the answer can be realized by computing the average natural (or pure) indirect effect:

$$NIE(X,Y) = E[Y_{x',Z_x}] - E[Y_{x'}]$$

which is the difference between the expected outcomes under no exposure when the mediator takes the value it would under exposure and non-exposure, respectively. We agree with the answer to the question (i.e., deactivating $X \to Y$) but not with its realization (i.e., deactivating $X \to Y$ by computing $NIE(X,Y)$), because the reference value x' affects the outcome in this realization of the answer. This is problematic because it means that the direct path $X \to Y$ is not really deactivated and, moreover, the answer depends on the level chosen as reference. In other words, this realization of the answer does not really correspond to the no-questioning policy being evaluated. The problems just discussed are shared by other classical measures of indirect effect, such as the average total indirect effect. However, this does not mean that these measures should be abandoned. Quite the opposite. They



(a) Original (b) G (c) H

Figure 1: Causal graphs in Sections 1 and 2.

are informative when the reference value is clear from the context. For instance, if women suspect being discriminated by the hirer, then they may want to know if the probability of a woman getting hired would remain unchanged had she a man's qualifications. This is measured by $NIE(X,Y)$ with reference value x' set to "woman". In summary, the existing measures of indirect effect are suitable in some cases and unsuitable in others. In this paper, we propose a new measure that does not require selecting a reference value.

An alternative to $NIE(X,Y)$ is the average interventional indirect effect developed by Geneletti (2007) (see also the work by VanderWeele et al. (2014)):

$$IIE(X,Y) = E[Y_{x',\mathcal{Z}_x}] - E[Y_{x',\mathcal{Z}_{x'}}]$$

where $\mathcal{Z}_x$ denotes a draw from the distribution of Z_x. This distribution may be estimated from a randomized controlled trial (RCT). Likewise for $\mathcal{Z}_{x'}$. Thus, the interventions on the mediator in $IIE(X,Y)$ are conceivable, while those in $NIE(X,Y)$ are not since the individual-specific Z_x and $Z_{x'}$ are never observed. Although $NIE(X,Y)$ and $IIE(X,Y)$ do not coincide in general, they do coincide for the causal graph in Figure 1(*a*) (VanderWeele et al., 2014). Therefore, the discussion in the previous paragraph also applies to $IIE(X,Y)$.

More recently, Fulcher et al. (2020) have introduced the population intervention indirect effect to measure the indirect effect of X on Y through the mediator Z:

$$PIIE(x') = E[Y_{X,Z_X}] - E[Y_{X,Z_{x'}}]$$

which is the difference between the expected outcomes when the exposure and mediator take natural (observed) values and when the exposure takes natural value but the mediator takes the value it would under no exposure. Therefore, this measure is suitable when the exposure is harmful (e.g., smoking) and, thus, one may be more interested in elucidating the effect (e.g., disease prevalence) of eliminating the exposure rather than in contrasting the effects of exposure and non-exposure. The latter is considered irrelevant, because it is inconceivable that everyone will be exposed. In this paper, though, we are interested in the latter because it may be informative even when the interventions are inconceivable. For instance, the two interventions being contrasted in the gender discrimination example above (everyone is male and everyone is female) are both inconceivable, but their contrast is instrumental to decide whether the no-questioning policy should be introduced or not, as argued by Pearl (2001, 2009) (see also the previous paragraph).

The rest of the paper is organized as follows. We present our new measure of indirect effect in Section 2. We also present a new measure of direct effect. We illustrate them with an example. Moreover, we show that they can partially handle unmeasured treatment-outcome confounding, and bound long-term effects by combining experimental

and observational data. We also introduce the concepts of indirect benefit and harm (i.e., through a mediator), and use our new measure to quantify them. Finally, we close with some discussion in Section 3.

2. Alternative Measures

Consider again the causal graph in Figure 1(*a*), which we call the original causal graph. We assume that the direct path $X \to Y$ is actually mediated by an unmeasured random variable U that is left unmodelled. This arguably holds for most direct paths. In the example above, U may represent the hirer's predisposition to hire the applicant. However, the identity of U is irrelevant. Let G denote the causal graph in Figure 1(*b*), i.e. the original causal graph refined with the addition of U. Now, deactivating the direct path $X \to Y$ in the original causal graph can be achieved by adjusting for U in G, i.e. $\sum_u E[Y|x,u]p(u)$. Unfortunately, U is unmeasured. We propose an alternative way of deactivating $X \to Y$. Let H denote the causal graph in Figure 1(*c*), i.e. the result of reversing the edge $X \to U$ in G. The average total effect of X on Y in H can be computed by the front-door criterion (Pearl, 2009):

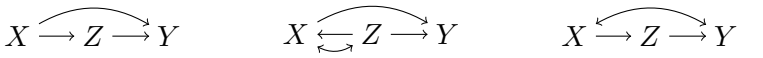
$$TE(X,Y) = E[Y_x] - E[Y_{x'}] = \sum_z p(z|x) \sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}) - \sum_z p(z|x') \sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}). \quad (1)$$

Note that every probability distribution that is representable by G is representable by H and vice versa (Pearl, 2009). Then, the right-hand side of the second equality in the equation above gives the same result whether it is evaluated in G or H. If we evaluate it in H, then it corresponds to the part of association between X and Y that is attributable to the path $X \to Z \to Y$. If we evaluate it in G, then it corresponds to the part of $TE(X,Y)$ in G that is attributable to the path $X \to Z \to Y$, because $TE(X,Y)$ in G equals the association between X and Y, since G has only directed paths from X to Y. Thus, the right-hand side of the second equality in the equation above corresponds to the part of $TE(X,Y)$ in the original causal graph that is attributable to the path $X \to Z \to Y$, thereby deactivating the direct path $X \to Y$. We therefore propose to use the right-hand side of the second equality in the equation above as a measure of the indirect effect of X on Y in the original causal graph:

$$IE(X,Y) = \sum_z p(z|x) \sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}) - \sum_z p(z|x') \sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}).$$

$IE(X,Y)$ only considers the path $X \to Z \to Y$ in the original causal graph to propagate the value of X. This is unlike $NIE(X,Y)$, which considers both paths from X to Y: The path $X \to Y$ propagates the value $X = x'$, whereas the path $X \to Z \to Y$ propagates the value that Z takes under $X = x$ and $X = x'$. As shown in Section 2.2, provided that Z is binary, $IE(X,Y)$ can be written as $TE(X,Z) \cdot TE(Z,Y)$, which some may find natural. It moreover coincides with the indirect effect in linear structural equation models.

An alternative way of motivating $IE(X,Y)$ is by interpreting the deactivation of the direct path $X \to Y$ in the original causal graph as hypothesizing that the domain under study can be represented by a causal graph that is equal to the original one save for the lack of the direct path $X \to Y$. Instead, the hypothesized graph has an edge $X \leftrightarrow Y$ representing the potential existence of an unmeasured treatment-outcome confounder. $IE(X,Y)$ in the original causal graph corresponds to $TE(X,Y)$ in the hypothesized one, which can be computed as in Equation 1.



(a) Original (b) Alternative 1 (c) Alternative 2

Figure 2: Causal graphs in Example 1.

Miles (2023) argues that an indirect effect measure should satisfy the following criterion in order to have a true mediational interpretation: The measure should be zero/non-negative/non-positive when the mediated effect is zero/non-negative/non-positive for all the individuals in the population. $NIE(X,Y)$ and $IIE(X,Y)$ satisfy the criterion.[1] $IE(X,Y)$ also satisfies the criterion because, as discussed above, it corresponds to the average of individual-level total effects in some causal graph.

Likewise, we propose to measure the direct effect of X on Y as the part of $TE(X,Y)$ in the original causal graph that remains after deactivating the path $X \to Z \to Y$. This is achieved by simply adjusting for Z:

$$DE(X,Y) = \sum_z E[Y|x,z]p(z) - \sum_z E[Y|x',z]p(z).$$

For the same reasons as above, this is unlike the measure proposed by Pearl (2001, 2009), namely the average natural (or pure) direct effect $NDE(X,Y) = E[Y_{x,Z_{x'}}] - E[Y_{x'}]$.

Finally, note that $NDE(X,Y)$ and $NIE(X,Y)$ can be computed from the observed data distribution $p(X,Z,Y)$ (Pearl, 2001, 2009). This is also true for $DE(X,Y)$ and $IE(X,Y)$. Like the sum of $NDE(X,Y)$ and $NIE(X,Y)$, the sum of $DE(X,Y)$ and $IE(X,Y)$ does not equal $TE(X,Y)$ in the original causal graph, due to interactions in the outcome model (Pearl, 2001, 2009; VanderWeele, 2013a, 2014). The discussion in this section also apply if, instead of $NDE(X,Y)$ and $NIE(X,Y)$, we consider the average total direct effect $TDE(X,Y) = E[Y_x] - E[Y_{x',Z_x}]$ and the average total indirect effect $TIE(X,Y) = E[Y_x] - E[Y_{x,Z_{x'}}]$, or the average controlled direct effect $CDE(X,Y) = E[Y_{x,\dot{z}}] - E[Y_{x',\dot{z}}]$ with $\dot{z} \in \{z, z'\}$.

We illustrate our measures $IE(X,Y)$ and $DE(X,Y)$ with the following example.[2]

Example 1 *Consider the following example studied by Pearl (2012). Figure 2(a) depicts the causal graph studied, hereinafter referred to as the original causal graph. In it, X represents a drug treatment, Z the presence of a certain enzyme in a patient's blood, and Y recovery. Moreover, we have that*

$$p(z|x) = 0.75 \quad p(y|x,z) = 0.8 \quad p(y|x,z') = 0.4$$
$$p(z|x') = 0.4 \quad p(y|x',z) = 0.3 \quad p(y|x',z') = 0.2.$$

Pearl imagines a scenario where someone proposes developing a cheaper drug that is equal to the existing one except for the lack of effect on enzyme production. To evaluate the new drug's performance, he computes $TE(X,Y) = 0.46$ and $NDE(X,Y) = 0.32$,

1. $IIE(X,Y)$ does not satisfy the criterion in general, unless additional assumptions are made. However, it does satisfy the criterion for the causal graph under consideration, because it coincides with $NIE(X,Y)$ (VanderWeele et al., 2014).
2. `R` code for the examples in this paper can be found at `https://tinyurl.com/2s3bxmyu`.

and concludes that the new drug will reduce the probability of recovery by 30%, *i.e.* $1 - NDE(X,Y)/TE(X,Y) = 0.3$. *We can repeat the analysis using* $DE(X,Y)$ *instead of* $NDE(X,Y)$:

$$\begin{aligned} DE(X,Y) &= p(y|x,z)p(z) + p(y|x,z')p(z') - p(y|x',z)p(z) - p(y|x',z')p(z') \\ &= 0.8p(z) + 0.4(1-p(z)) - 0.3p(z) - 0.2(1-p(z)) \\ &= 0.2 + 0.3p(z) = 0.2 + 0.3[p(z|x)p(x) + p(z|x')p(x')] \\ &= 0.2 + 0.3[0.75p(x) + 0.4(1-p(x))] = 0.32 + 0.11p(x) \end{aligned}$$

which implies that $0.32 \leq DE(X,Y) \leq 0.43$. *An interval is returned because* $p(X)$ *is not given in the original example (it is not needed to compute* $NDE(X,Y)$ *or* $NIE(X,Y)$*). Therefore, we conclude that the new drug will reduce the probability of recovery by between* 7% *and* 30%, *depending on* $p(X)$.

*The new drug development scenario described above corresponds to the alternative causal graph in Figure 2(*b*). The edge* $X \leftarrow Z$ *represents that the presence of enzyme may have an effect on the patient taking the treatment, and* $X \leftrightarrow Z$ *represents the potential existence of an unmeasured confounder between them. The drug performance in this alternative causal graph is simply* $TE(X,Y)$, *which can be computed by adjusting for* Z, *and thus it coincides with* $DE(X,Y)$ *in the original causal graph. In other words, it is* $DE(X,Y)$ *rather than* $NDE(X,Y)$ *that should be used to answer the original question. Note that* $DE(X,Y) = NDE(X,Y) = 0.32$ *if and only if* $p(x) = 0$, *i.e. everyone is untreated. This is no coincidence because* $NDE(X,Y)$ *in the original causal graph coincides with the average effect of the treatment among the untreated in the alternative graph (Ogburn and VanderWeele, 2012b),*[3] *rather than with* $TE(X,Y)$ *which is the correct answer to the original question.*

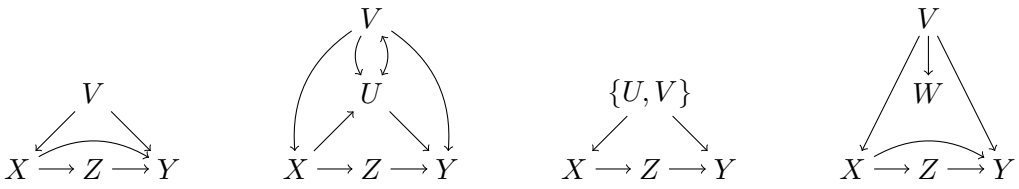
Pearl also imagines a scenario where someone proposes developing a cheaper drug that is equal to the existing one except for the lack of direct effect on recovery, i.e. it just stimulates enzyme production as much as the existing drug. To evaluate the new drug's performance, he computes $TE(X,Y) = 0.46$ *and* $NIE(X,Y) = 0.04$, *and concludes that the new drug will reduce the probability of recovery by* 91%, *i.e.* $1 - NIE(X,Y)/TE(X,Y) = 0.91$.[4] *We can repeat the analysis using* $IE(X,Y)$ *instead of* $NIE(X,Y)$:

$$\begin{aligned} IE(X,Y) &= p(z|x)[p(y|x,z)p(x) + p(y|x',z)p(x')] + p(z'|x)[p(y|x,z')p(x) + p(y|x',z')p(x')] \\ &\quad - p(z|x')[p(y|x,z)p(x) + p(y|x',z)p(x')] - p(z'|x')[p(y|x,z')p(x) + p(y|x',z')p(x')] \\ &= 0.75[0.8p(x) + 0.3(1-p(x))] + 0.25[0.4p(x) + 0.2(1-p(x))] \\ &\quad - 0.4[0.8p(x) + 0.3(1-p(x))] - 0.6[0.4p(x) + 0.2(1-p(x))] = 0.04 + 0.11p(x) \end{aligned}$$

which implies that $0.04 \leq IE(X,Y) \leq 0.15$. *Therefore, we conclude that the new drug will reduce the probability of recovery by between* 67% *and* 91%, *depending on* $p(X)$.

*The latest new drug development scenario corresponds to the alternative causal graph in Figure 2(*c*). The edge* $X \leftrightarrow Y$ *represents the potential existence of an unmeasured treatment-outcome confounder. The drug performance in this alternative causal graph is simply* $TE(X,Y)$, *which can be computed by the front-door criterion, and thus it coincides*

3. Ogburn and VanderWeele (2012b) prove the equivalence when $X \leftarrow Z$, but the proof also applies when $X \leftrightarrow Z$.
4. The small disagreements with the results by Pearl (2012) are due to rounding.



(a) Extended (b) Add U (c) Group U and V (d) Proxy W

Figure 3: Causal graphs in Section 2.1.

with $IE(X,Y)$ in the original causal graph. In other words, it is $IE(X,Y)$ rather than $NIE(X,Y)$ that should be used to answer the original question. Note that $IE(X,Y) = NIE(X,Y) = 0.04$ if and only if $p(x) = 0$. Again, this is no coincidence because $NIE(X,Y)$ in the original causal graph coincides with the average effect of the treatment among the untreated in the alternative graph (Pearl, 2001; Shpitser and Pearl, 2009), rather than with $TE(X,Y)$ which is the correct answer to the original question.

2.1. Unmeasured Confounding

In this section, we extend the original causal graph in the previous section with an unmeasured treatment-outcome confounder V. See Figure 3(*a*). Now, neither $NDE(X,Y)$ nor $NIE(X,Y)$ nor their total, controlled and interventional counterparts are identifiable from the observed data distribution $p(X,Z,Y)$ (Pearl, 2001, 2009; VanderWeele et al., 2014). However, $IE(X,Y)$ can be computed pretty much like before. First, we add the unmeasured mediator U. See Figure 3(*b*). Then, we group U and V. See Figure 3(*c*). Note that every probability distribution that is representable by the graph in Figure 3(*b*) is representable by the graph in Figure 3(*c*), since all the independencies entailed by the latter hold in the former. Finally, we apply the front-door criterion.

Like $NDE(X,Y)$ and its total, controlled and interventional counterparts, $DE(X,Y)$ is not identifiable from the observed data distribution $p(X,Z,Y)$ in the extended causal graph under consideration. However, it may be bounded if V is binary and a binary proxy W of V is measured. The causal graph under consideration is then the one in Figure 3(*d*). In the literature, there are many cautionary tales about the bias that adjusting for the proxy of an unmeasured confounder introduces to the estimation of a causal effect (Austin and Brunner, 2004; Altman and Royston, 2006; Chen et al., 2007a). For instance, Brenner (1997) constructs an example where adjusting for the proxy is worse than not adjusting at all. However, there are conditions under which the opposite is true (Gabriel et al., 2022; Ogburn and VanderWeele, 2012a; Peña, 2020; Sjölander et al., 2022). We use some of these conditions here.

Recall that $DE(X,Y)$ is the part of $TE(X,Y)$ in the causal graph that remains after deactivating the path $X \to Z \to Y$. This is achieved by simply adjusting for Z:

$$DE(X,Y) = \sum_z TE(X,Y|z)p(z)$$

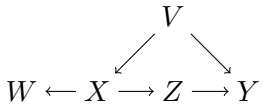


Figure 4: Causal graph in Section 2.2.

where $TE(X, Y|z)$ is the average total effect of X on Y in the stratum $Z = z$. Let us define the observed or partially adjusted average total effect of X on Y in the stratum $Z = z$ as

$$TE_{obs}(X, Y|z) = \sum_w E[Y|x, z, w]p(w|z) - \sum_w E[Y|x', z, w]p(w|z).$$

Note that it can be computed from the observed data distribution $p(X, W, Z, Y)$. Rephrasing the results by Ogburn and VanderWeele (2012a) and Peña (2020) to our scenario, if $E[Y|\dot{x}, \dot{z}, W]$ and $E[X|\dot{z}, W]$ are one nonincreasing and the other nondecreasing in W for all $\dot{x} \in \{x, x'\}$ and $\dot{z} \in \{z, z'\}$, then $TE(X, Y|\dot{z}) \geq TE_{obs}(X, Y|\dot{z})$ for all $\dot{z} \in \{z, z'\}$. On the other hand, if $E[Y|\dot{x}, \dot{z}, W]$ and $E[X|\dot{z}, W]$ are both nonincreasing or both nondecreasing in W for all $\dot{x} \in \{x, x'\}$ and $\dot{z} \in \{z, z'\}$, then $TE_{obs}(X, Y|\dot{z}) \geq TE(X, Y|\dot{z})$ for all $\dot{z} \in \{z, z'\}$. Note that the antecedents of these rules are testable from the observed data distribution $p(X, W, Z, Y)$. Not in all but in many cases, these rules enable us to bound $DE(X, Y)$ and even determine its sign. Specifically,

$$DE(X, Y) \cdot (2 \cdot \mathbf{1}_{\neq} - 1) \geq [\sum_z TE_{obs}(X, Y|z)p(z)] \cdot (2 \cdot \mathbf{1}_{\neq} - 1)$$

where $\mathbf{1}_{\neq}$ is 1 (respectively, 0) if $E[Y|\dot{x}, \dot{z}, W]$ and $E[X|\dot{z}, W]$ are one nonincreasing and the other nondecreasing (respectively, both nonincreasing or both nondecreasing) in W for all $\dot{x} \in \{x, x'\}$ and $\dot{z} \in \{z, z'\}$.

2.2. Long-Term Effects

This section addresses a problem of RCTs, namely long-time effect estimation from typically short-lived trials. Consider the causal graph in Figure 4, where X and V are unmeasured. We assume that the mediator Z is a short-term effect of the treatment X, whereas Y is a long-term effect of X. RCTs are typically short-lived due to cost considerations and, thus, they are typically conducted to estimate short-term effects but not longer ones. Observational data, on the other hand, is much cheaper to obtain and, thus, they may include long-term outcome observations. Unfortunately, observational data is typically subject to unmeasured confounding, and mismeasurements due to self-reporting. Therefore, we assume that a RCT was conducted to estimate $TE(X, Z)$, but not $TE(Z, Y)$ or $TE(X, Y)$. We also assume that the probability distribution $p(W, Z, Y)$ was estimated from observational data, where W represents the self-reported treatment, which may differ from the actual unmeasured treatment X. Our goal is computing $TE(X, Y)$. Unfortunately, this cannot be done from the information available. However, the fact that $TE(X, Y) = IE(X, Y)$ implies, as we show below, that $TE(X, Y)$ can be bounded sometimes.

Our setup above is similar to the one by Athey et al. (2019) with the differences that they assume no unmeasured confounding, and that neither the true nor the self-reported

treatment is available in their observational data. Our setup is also close to the ones by Athey et al. (2020), Ghassami et al. (2022), Imbens et al. (2022) and Van Goffrier et al. (2023) with the differences that their unmeasured confounders affect both the short-term and long-term outcomes, and that the true treatment is available in their observational data. Finally, Appendix A discusses how our setup fits within the literature on surrogate endpoints.

Provided that Z is binary, we have that

$$\begin{aligned}
IE(X,Y) &= p(z|x)\sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}) + p(z'|x)\sum_{\dot{x}} E[Y|\dot{x},z']p(\dot{x}) \\
&\quad - p(z|x')\sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}) - p(z'|x')\sum_{\dot{x}} E[Y|\dot{x},z']p(\dot{x}) \\
&= [p(z|x) - p(z|x')][\sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x})] + [p(z'|x) - p(z'|x')][\sum_{\dot{x}} E[Y|\dot{x},z']p(\dot{x})] \\
&= [p(z|x) - p(z|x')][\sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x})] + [-p(z|x) + p(z|x')][\sum_{\dot{x}} E[Y|\dot{x},z']p(\dot{x})] \\
&= [p(z|x) - p(z|x')][\sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x}) - \sum_{\dot{x}} E[Y|\dot{x},z']p(\dot{x})] \\
&= [E[Z_x] - E[Z_{x'}]][E[Y_z] - E[Y_{z'}]] = TE(X,Z)\cdot TE(Z,Y).
\end{aligned}$$

Let us define the observed or partially adjusted average total effect of Z on Y as

$$TE_{obs}(Z,Y) = \sum_{w} E[Y|z,w]p(w) - \sum_{w} E[Y|z',w]p(w).$$

Note that it can be computed from the observed data distribution $p(W,Z,Y)$. If $E[Y|\dot{z},W]$ and $E[Z|W]$ are one nonincreasing and the other nondecreasing in W for all $\dot{z} \in \{z,z'\}$, then $TE(Z,Y) \geq TE_{obs}(Z,Y)$ (Ogburn and VanderWeele, 2012a; Peña, 2020). On the other hand, if $E[Y|\dot{z},W]$ and $E[Z|W]$ are both nonincreasing or both nondecreasing in W for all $\dot{z} \in \{z,z'\}$, then $TE_{obs}(Z,Y) \geq TE(Z,Y)$ (Ogburn and VanderWeele, 2012a; Peña, 2020).[5] Note that the antecedents of these rules are testable from the observed data distribution $p(W,Z,Y)$. Not in all but in many cases, these rules together with the knowledge of $TE(X,Z)$ enable us to bound $IE(X,Y)$ and even determine its sign. Specifically,

$$IE(X,Y)\cdot(2\cdot\mathbf{1}_{\neq}-1)\cdot(2\cdot\mathbf{1}_{\geq}-1) \geq TE(X,Z)\cdot TE_{obs}(Z,Y)\cdot(2\cdot\mathbf{1}_{\neq}-1)\cdot(2\cdot\mathbf{1}_{\geq}-1) \quad (2)$$

where $\mathbf{1}_{\neq}$ is 1 (respectively, 0) if $E[Y|\dot{z},W]$ and $E[Z|W]$ are one nonincreasing and the other nondecreasing (respectively, both nonincreasing or both nondecreasing) in W for all $\dot{z} \in \{z,z'\}$, and $\mathbf{1}_{\geq}$ is 1 if $TE(X,Z) \geq 0$ and 0 otherwise.

Finally, note that Equation 2 also holds if we add the edge $X \to Y$ to the causal graph under study. To see it, simply pre-process the graph as we did at the beginning of Section 2.1.

5. In the proofs of these results, X is a parent of V. The results also hold when X is a child of V, since (i) every probability distribution that is representable when X is a child of V is also representable when X is a parent of V and vice versa (Pearl, 2009), and (ii) $TE(Z,Y)$ and $TE_{obs}(Z,Y)$ give each the same result in both cases.

2.3. Indirect Benefit and Harm

This section introduces the concepts of indirect benefit and harm (i.e., through a mediator), and shows how to measure them. Let X and Y denote an exposure and its outcome, respectively. Let X and Y be binary taking values in $\{x, x'\}$ and $\{y, y'\}$. Let Y_x and $Y_{x'}$ denote the counterfactual outcome when the exposure is set to level $X = x$ and $X = x'$. Let y_x, y'_x, $y_{x'}$ and $y'_{x'}$ denote the events $Y_x = y$, $Y_x = y'$, $Y_{x'} = y$ and $Y_{x'} = y'$. For instance, let X represent whether a patient gets treated or not for a deadly disease, and Y represent whether she survives it or not. Individual patients can be classified into immune (they survive whether they are treated or not, i.e. $y_x \wedge y_{x'}$), doomed (they die whether they are treated or not, i.e. $y'_x \wedge y'_{x'}$), benefited (they survive if and only if treated, i.e. $y_x \wedge y'_{x'}$), and harmed (they die if and only if treated, i.e. $y'_x \wedge y_{x'}$).

In general, the average treatment effect (ATE) estimated from a RCT does not inform about the probability of benefit (or of any of the other response types, i.e. harm, immunity, and doom). However, it may do it under certain conditions. For instance,

$$\begin{aligned} ATE &= p(y_x) - p(y_{x'}) = p(y_x, y_{x'}) + p(y_x, y'_{x'}) - [p(y_x, y_{x'}) + p(y'_x, y_{x'})] \\ &= p(y_x, y'_{x'}) - p(y'_x, y_{x'}) = p(\text{benefit}) - p(\text{harm}) \end{aligned} \tag{3}$$

and thus $p(\text{benefit}) = ATE$ if $p(\text{harm}) = 0$ (a.k.a. monotonicity (Pearl, 2009)). Mueller and Pearl (2023) derive necessary and sufficient conditions to determine from observational and experimental data if monotonicity holds. We derive below similar conditions for non-immunity, i.e. $p(\text{immunity}) = p(y_x, y_{x'}) = 0$. These are interesting because under non-monotonicity, they turn an RCT informative about the probabilities of benefit and harm. To see it, consider

$$ATE = p(y_x) - p(y_{x'})$$

where the terms on the right-hand side of the equation are estimated from an RCT. Moreover,

$$p(y_x) = p(y_x, y_{x'}) + p(y_x, y'_{x'}) = p(\text{immunity}) + p(\text{benefit}) \tag{4}$$

and

$$p(y_{x'}) = p(y_x, y_{x'}) + p(y'_x, y_{x'}) = p(\text{immunity}) + p(\text{harm}) \tag{5}$$

and thus $p(\text{benefit}) = p(y_x)$ and $p(\text{harm}) = p(y_{x'})$ if $p(\text{immunity}) = 0$.

To derive necessary and sufficient conditions for non-immunity, consider first the bounds of $p(\text{benefit})$ derived by Tian and Pearl (2000):

$$\max \left\{ \begin{array}{c} 0, \\ p(y_x) - p(y_{x'}), \\ p(y) - p(y_{x'}), \\ p(y_x) - p(y) \end{array} \right\} \leq p(\text{benefit}) \leq \min \left\{ \begin{array}{c} p(y_x), \\ p(y'_{x'}), \\ p(x, y) + p(x', y'), \\ p(y_x) - p(y_{x'}) + p(x, y') + p(x', y) \end{array} \right\}. \tag{6}$$

Then, combining Equations 4 or 5 with 6 gives

$$\max \left\{ \begin{array}{c} 0, \\ p(y_x) - p(y'_{x'}), \\ p(y_x) - p(x, y) - p(x', y'), \\ p(y_{x'}) - p(x, y') - p(x', y) \end{array} \right\} \leq p(\text{immunity}) \leq \min \left\{ \begin{array}{c} p(y_x), \\ p(y_{x'}), \\ p(y_x) - p(y) + p(y_{x'}), \\ p(y) \end{array} \right\}. \tag{7}$$

A sufficient condition for $p(\text{immunity}) = 0$ to hold is that some argument to the min function in Equation 7 is equal to 0, that is

$$p(y_x) = 0 \text{ or } p(y_{x'}) = 0 \text{ or } p(y_x) + p(y_{x'}) = p(y) \text{ or } p(y) = 0. \tag{8}$$

Likewise, a necessary condition for $p(\text{immunity}) = 0$ to hold is that all the arguments to the max function are non-positive, that is

$$p(y_x) + p(y_{x'}) \leq 1 \text{ and } p(y_x) \leq p(x, y) + p(x', y') \text{ and } p(y_{x'}) \leq p(x, y') + p(x', y). \tag{9}$$

The previous conditions for non-immunity can be relaxed to allow certain degree of immunity (e.g., based on expert knowledge), making them more applicable in practice as we show below. Specifically, a sufficient condition for $p(\text{immunity}) \leq \epsilon$ to hold is

$$p(y_x) \leq \epsilon \text{ or } p(y_{x'}) \leq \epsilon \text{ or } p(y_x) + p(y_{x'}) \leq p(y) + \epsilon \text{ or } p(y) \leq \epsilon. \tag{10}$$

Likewise, a necessary condition for $p(\text{immunity}) \leq \epsilon$ to hold is

$$p(y_x) + p(y_{x'}) \leq 1 + \epsilon \text{ and } p(y_x) \leq p(x, y) + p(x', y') + \epsilon \text{ and } p(y_{x'}) \leq p(x, y') + p(x', y) + \epsilon. \tag{11}$$

These conditions for ϵ-bounded immunity can now be used to narrow the bounds on $p(\text{benefit})$ in Equation 6. Specifically, if $p(\text{immunity}) \leq \epsilon$ then Equation 4 gives

$$p(y_x) - \epsilon \leq p(\text{benefit}) \leq p(y_x).$$

Incorporating this into Equation 6 gives

$$\max \left\{ \begin{array}{c} 0, \\ p(y_x) - p(y_{x'}), \\ p(y) - p(y_{x'}), \\ p(y_x) - p(y), \\ p(y_x) - \epsilon \end{array} \right\} \leq p(\text{benefit}) \leq \min \left\{ \begin{array}{c} p(y_x), \\ p(y'_{x'}), \\ p(x, y) + p(x', y'), \\ p(y_x) - p(y_{x'}) + p(x, y') + p(x', y) \end{array} \right\} \tag{12}$$

which returns a tighter lower bound than Equation 6 if $\epsilon < \min(p(y_{x'}), p(y))$. Although the value of ϵ is typically determined from expert knowledge and not from data, the experimental and observational data available do restrict the values that are valid, as indicated by Equation 11. Alternatively, ϵ can take any value as long as the lower bound is not greater than the upper bound in Equation 12. Moreover, $p(\text{harm})$ can likewise be bounded by simply swapping x and x' in Equation 12.

We illustrate our improved bounds with the following example.

Example 2 *A pharmaceutical company wants to market their drug to cure a disease by claiming that no one is immune. The RCT they conducted for the drug approval yielded the following:*

$$p(y_x) = 0.76 \qquad\qquad p(y_{x'}) = 0.31$$

which correspond to the following unknown data generation model:

$$\begin{array}{llll} p(u) = 0.3 & p(x|u) = 0.2 & p(y|x,u) = 0.9 & p(y|x,u') = 0.7 \\ & p(x|u') = 0.9 & p(y|x',u) = 0.8 & p(y|x',u') = 0.1. \end{array}$$

Therefore, the necessary condition for non-immunity in Equation 9 does not hold, and thus the company is not entitled to make the claim they intended to make. The company changes strategy and now wishes to market their drug as having a minimum of 50 % efficacy, i.e. benefit. To do so, they first conduct an observational study that yields the following:

$$p(x,y) = 0.5 \qquad p(x,y') = 0.2 \qquad p(x',y) = 0.2 \qquad p(x',y') = 0.1.$$

Then, they apply Equation 6 to the RCT and observational results yielding $0.45 \leq p(\textit{benefit}) \leq 0.61$. *Again, the company cannot proceed with their marketing strategy. A few months later, a research publication reports that no more than 25 % of the population is immune. The company realizes that this value is compatible with their RCT and observational results, by checking the necessary condition for ϵ-bounded immunity in Equation 11. More importantly, the company realizes that Equation 12 with $\epsilon = 0.25$ allows to conclude that* $0.51 \leq p(\textit{benefit}) \leq 0.61$, *and thus they can resume their latest marketing strategy.*

So far in this section, the causal graph of the domain under study was unknown. Now, we consider again the causal graph in Figure 1(*a*), and we show how to compute the probabilities of benefit and harm mediated by Z.[6] Recall that we defined G and H as the causal graphs in Figures 1(*b*) and 1(*c*) respectively, and we noted that they represent different data generation mechanisms but the same probability distributions over X, Y and Z. Therefore, the mechanisms agree on observational probabilities but may disagree on counterfactual probabilities. We use $p()$ to denote observational probabilities obtained from either mechanism, and $q()$ to denote counterfactual probabilities obtained from the mechanism corresponding to H. The probabilities of benefit and harm of X on Y mediated by Z in G and thus in the original causal graph (henceforth indirect benefit and harm, or IB and IH) can be computed by applying Equation 4 to H. That is,

$$IB = q(\text{benefit}) = q(y_x) = \sum_z p(z|x) \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x})$$

where the second equality holds if $q(\text{immunity}) = 0$, and the third is due to the front-door criterion on H. Likewise for IH simply replacing x by x' due to Equation 5. Applying Equation 7 to H yields necessary and sufficient conditions for $q(\text{immunity}) = 0$. That is,

$$\begin{aligned} &\sum_z p(z|x) \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) = 0 \text{ or } \sum_z p(z|x') \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) = 0 \text{ or} \\ &\sum_z [p(z|x) + p(z|x')] \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) = p(y) \text{ or } p(y) = 0 \end{aligned} \tag{13}$$

is a sufficient condition, whereas

$$\begin{aligned} &\sum_z [p(z|x) + p(z|x')] \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) \leq 1 \text{ and } \sum_z p(z|x) \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) \leq p(x,y) + p(x',y') \\ &\text{and } \sum_z p(z|x') \sum_{\dot{x}} p(y|\dot{x},z)p(\dot{x}) \leq p(x,y') + p(x',y) \end{aligned} \tag{14}$$

6. Note that for this causal graph, $p(y_x) = p(y|x)$ and $p(y_{x'}) = p(y|x')$ and thus, $p(y_x)$ and $p(y_{x'})$ can be estimated from observational data and thus, no RCT is actually required for our previous results in this section.

is a necessary condition. Necessary and sufficient conditions for ϵ-bounded immunity on H (i.e., $q(\text{immunity}) \leq \epsilon$) can be obtained much like in Equations 10 and 11. That is, it suffices to add ϵ to the right-hand sides of the conditions above and replace $=$ with $\leq$. Finally, we can adapt accordingly Equation 12 to obtain ϵ-bounds on IB and IH. Note that the analysis of indirect benefit and harm presented here does not require an RCT, i.e. all the expressions involved can be estimated from just observational data.

We illustrate the results above with the following example.

Example 3 *Consider again the model in Example 1, which is borrowed from Pearl (2012). Since $p(x)$ is not given in the original example, we take $p(x) = 0.6$.*

Pearl imagines a scenario where the pharmaceutical company plans to develop a cheaper drug that is equal to the existing one except for the lack of direct effect on recovery, i.e. it just stimulates enzyme production as much as the existing drug. Then, the probability of benefit of the planned drug is the probability of benefit of the existing drug that is mediated by the enzyme. The company wants to market their drugs by claiming that no one is immune. The sufficient conditions for non-immunity in Equations 8 and 13 do not hold for the drugs. However, while the existing drug satisfies the necessary condition for non-immunity in Equation 9, the planned drug does not satisfy the corresponding condition in Equation 14. Then, the company should either abandon their marketing strategy or abandon the plan to develop the new drug and instead focus on confirming non-immunity for the existing drug.

3. Discussion

We have proposed new measures of direct and indirect effects. They are based on contrasting the effects of exposure and non-exposure and they do not require selecting a reference value. This makes them unlike the existing measures in the literature and, thus, suitable in cases where the existing ones are unsuitable. The opposite is also true. The new measures assume that the direct path from expose to outcome is mediated by an unmeasured random variable. Its identity is irrelevant. This arguably holds for most direct paths.

When there is unmeasured treatment-outcome confounding, we have shown that the new measure of indirect effect still applies, whereas the new measure of direct effect can be bounded in some cases. This also sets them apart from the existing measures. Moreover, we have shown how the new measure of indirect effect can be used to sometimes bound long-term effects by combining experimental and observational data. Finally, we have used the new measure to quantify the indirect benefit and harm. To do so, we have first studied the probability of immunity and derived necessary and sufficient conditions for non-immunity and ϵ-bounded immunity. We have also shown how these results tighten the existing bounds of the probabilities of benefit and harm. Moreover, Appendix B presents a method for sensitivity analysis of the probability of immunity under unmeasured confounding.

Note that some of our results require some random variables being binary, specifically the mediator Z and the unmeasured confounder V. We plan to investigate extensions to discrete random variables of arbitrary cardinalities. Then, our results would hold for larger causal graphs than the ones considered in this paper, as Z and V could be sets of random variables. In other words, our results would hold for larger causal graphs as long as they can be projected onto the ones considered in this paper by grouping the mediators and unmeasured confounders into Z and V, respectively.

References

D. G. Altman and P. Royston. The Cost of Dichotomising Continuous Variables. *BMJ*, 332:1080, 2006.

S. Athey, R. Chetty, G. W. Imbens, and H. Kang. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. Technical Report 26463, National Bureau of Economic Research, 2019.

S. Athey, R. Chetty, and G. W. Imbens. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. *arXiv:2006.09676 [stat.ME]*, 2020.

P. C. Austin and L. J. Brunner. Inflation of the Type I Error Rate when a Continuous Confounding Variable is Categorized in Logistic Regression Analyses. *Statistics in medicine*, 23:1159–1178, 2004.

H. Brenner. A Potential Pitfall in Control of Covariates in Epidemiologic Studies. *Epidemiology*, 9:68–71, 1997.

H. Chen, P. Cohen, and S. Chen. Biased Odds Ratios from Dichotomization of Age. *Statistics in medicine*, 26:3487–3497, 2007a.

H. Chen, Z. Geng, and J. Jia. Criteria for Surrogate End Points. *Journal of the Royal Statistical Society B*, 69:919–932, 2007b.

I. R. Fulcher, I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen. Robust Inference on Population Indirect Causal Effects: The Generalized Front Door Criterion. *Journal of the Royal Statistical Society Series B*, 82:199–214, 2020.

E. E. Gabriel, J. M. Peña, and A. Sjölander. Bias Attenuation Results for Dichotomization of a Continuous Confounder. *Journal of Causal Inference*, 10:515–526, 2022.

S. Geneletti. Identifying Direct and Indirect Effects in a Non-Counterfactual Framework. *Journal of the Royal Statistical Society Series B*, 69:199–215, 2007.

A. Ghassami, D. Yang, A. Richardson, I. Shpitser, and E. Tchetgen Tchetgen. Combining Experimental and Observational Data for Identification and Estimation of Long-Term Causal Effects. *arXiv:2201.10743 [stat.ME]*, 2022.

P. B. Gilbert and M. G. Hudgens. Evaluating Candidate Principal Surrogate Endpoints. *Biometrics*, 64:1146–1154, 2008.

G. Imbens, N. Kallus, X. Mao, and Y. Wang. Long-term Causal Inference Under Persistent Confounding via Data Combination. *arXiv:2202.07234 [stat.ME]*, 2022.

C. Ju and Z. Geng. Criteria for Surrogate End Points Based on Causal Distributions. *Journal of the Royal Statistical Society B*, 72:129–142, 2010.

C. H. Miles. On the Causal Interpretation of Randomised Interventional Indirect Effects. *Journal of the Royal Statistical Society Series B*, 00:1–19, 2023.

S. Mueller and J. Pearl. Monotonicity: Detection, Refutation, and Ramification. *UCLA Cognitive Systems Laboratory, Technical Report (R-529)*, 2023.

E. L. Ogburn and T. J. VanderWeele. On the Nondifferential Misclassification of a Binary Confounder. *Epidemiology*, 23:433–439, 2012a.

E. L. Ogburn and T. J. VanderWeele. Analytic Results on the Bias Due to Nondifferential Misclassification of a Binary Mediator. *American Journal of Epidemiology*, 176:555–561, 2012b.

J. M. Peña. On the Monotonicity of a Nondifferentially Mismeasured Binary Confounder. *Journal of Causal Inference*, 8:150–163, 2020.

J. M. Peña. Bounding the Probabilities of Benefit and Harm Through Sensitivity Parameters and Proxies. *Journal of Causal Inference*, 11:20230012, 2023.

J. Pearl. Direct and Indirect Effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

J. Pearl. The Causal Mediation Formula - A Guide to the Assessment of Pathways and Mechanisms. *Prevention Science*, 13:426–436, 2012.

I. Shpitser and J. Pearl. Effects of Treatment on the Treated: Identification and Generalization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 514–521, 2009.

A. Sjölander, J. M. Peña, and E. E. Gabriel. Bias Results for Nondifferential Mismeasurement of a Binary Confounder. *Statistics & Probability Letters*, 186:109474, 2022.

J. Tian and J. Pearl. Probabilities of Causation: Bounds and Identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 2000.

G. Van Goffrier, L. Maystre, and C. M. Gilligan-Lee. Estimating Long-Term Causal Effects from Short-Term Experiments and Long-Term Observational Data with Unobserved Confounding. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning*, 2023.

T. J. VanderWeele. A Three-Way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects. *Epidemiology*, 24:224–232, 2013a.

T. J. VanderWeele. Surrogate Measures and Consistent Surrogates. *Biometrics*, 69:561–581, 2013b.

T. J. VanderWeele. A Unification of Mediation and Interaction: A Four-Way Decomposition. *Epidemiology*, 25:749–761, 2014.

T. J. VanderWeele, S. Vansteelandt, and J. M. Robins. Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology*, 25:300–306, 2014.

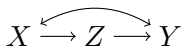


Figure 5: Causal graph in Appendix B.

Z. Wu, P. He, and Z. Geng. Sufficient Conditions for Concluding Surrogacy Based on Observed Data. *Statistics in Medicine*, 30:2422–2434, 2011.

Appendix A. Consistent Surrogates

As mentioned in Section 2.2, RCTs do not typically measure the true outcome of interest and instead, they measure a surrogate of it. The surrogate is usually a mediator of the outcome of interest, but it suffices with the surrogate being predictive of the outcome of interest (VanderWeele, 2013b). Moreover, it is desirable that the surrogate is consistent, i.e. the sign of the ATE of the treatment on the surrogate is predictive of the sign of the ATE of the treatment on the outcome of interest. Otherwise, the so-called surrogate paradox occurs (Chen et al., 2007b; Ju and Geng, 2010; VanderWeele, 2013b). The existing criteria for selecting or validating surrogates can be divided into empirical (i.e., testable from the data available) and a priori (i.e., based on domain knowledge and thus untestable). The empirical criteria suffer from the surrogate paradox (VanderWeele, 2013b), with the only exception of those by Gilbert and Hudgens (2008) and Wu et al. (2011). On the other hand, the a priori criteria by Chen et al. (2007b), Ju and Geng (2010) and VanderWeele (2013b) avoid the surrogate paradox but they are difficult to use in practice.

In the setup studied in Section 2.2, Equation 2 can be seen as an empirical criterion for validating Z as a consistent surrogate whenever $TE(X,Z)\cdot TE_{obs}(Z,Y)\cdot(2\cdot\mathbf{1}_{\neq}-1)\cdot(2\cdot\mathbf{1}_{\geq}-1)$ is positive. Our setup differs from the setups considered by Gilbert and Hudgens (2008) and Wu et al. (2011) in that our observational data include unmeasured treatment-outcome confounding and self-reported treatment.

Appendix B. Sensitivity Analysis of Immunity under Confounding

Assume that we only have access to observational data, i.e. no RCT is available. Consider the causal graph in Figure 5, which includes potential unmeasured exposure-outcome confounding. Since

$$p(y_x) = \sum_z p(z|x) \sum_{\dot{x}} E[Y|\dot{x},z]p(\dot{x})$$

by the front-door criterion, we can proceed as in Equations 13 and 14 to derive necessary and sufficient conditions for non-immunity. Suppose now that Z is unmeasured or that the effect of X on Y is direct rather than mediated by Z. Then, $p(y_x)$ is unidentifiable from observational data (Pearl, 2009), and thus we cannot proceed as indicated. We therefore take an alternative approach to inform the analyst about the probability of immunity and thereby help her in decision making. In particular, we propose a sensitivity analysis method to bound the probability of immunity as a function of the observed data distribution and

some intuitive sensitivity parameters. Our method is an straightforward adaption of the method by Peña (2023), originally developed to bound the probabilities of benefit and harm.

Let U denote the unmeasured exposure-outcome confounders. For simplicity, we assume that all these confounders are categorical, but our results also hold for ordinal and continuous confounders.[7] For simplicity, we treat U as a categorical random variable whose levels are the Cartesian product of the levels of the elements in the original U.

Note that

$$p(y_x) = p(y_x|x)p(x) + p(y_x|x')p(x') = p(y|x)p(x) + p(y_x|x')p(x')$$

where the second equality follows from counterfactual consistency, i.e. $X = x \Rightarrow Y_x = Y$. Moreover,

$$p(y_x|x') = \sum_u p(y_x|x', u)p(u|x') = \sum_u p(y|x, u)p(u|x') \leq \max_u p(y|x, u)$$

where the second equality follows from $Y_x \perp X|U$ for all x, and counterfactual consistency. Likewise,

$$p(y_x|x') \geq \min_u p(y|x, u).$$

Now, let us define

$$M_x = \max_u p(y|x, u)$$

and

$$m_x = \min_u p(y|x, u)$$

and likewise $M_{x'}$ and $m_{x'}$. Then,

$$p(x, y) + p(x')m_x \leq p(y_x) \leq p(x, y) + p(x')M_x$$

and likewise

$$p(x', y) + p(x)m_{x'} \leq p(y_{x'}) \leq p(x', y) + p(x)M_{x'}.$$

These equations together with Equation 7 give

$$\max \left\{ \begin{array}{c} 0, \\ p(x')m_x + p(x)m_{x'} - p(y'), \\ p(x')m_x - p(x', y'), \\ p(x)m_{x'} - p(x, y') \end{array} \right\} \leq p(\text{immunity}) \leq \min \left\{ \begin{array}{c} p(x, y) + p(x')M_x, \\ p(x', y) + p(x)M_{x'}, \\ p(x')M_x + p(x)M_{x'}, \\ p(y) \end{array} \right\} \quad (15)$$

where m_x, M_x, $m_{x'}$ and $M_{x'}$ are sensitivity parameters. The possible regions for m_x and M_x are

$$0 \leq m_x \leq p(y|x) \leq M_x \leq 1 \quad (16)$$

and likewise for $m_{x'}$ and $M_{x'}$.

Our lower bound in Equation 15 is informative if and only if[8]

$$0 < p(x')m_x - p(x', y')$$

7. If U is continuous then sums/maxima/minimima over u should be replaced by integrals/suprema/infima.
8. Note that the second row in the maximum equals the third plus the fourth rows.

or

$$0 < p(x)m_{x'} - p(x, y').$$

Then, the informative regions for m_x and $m_{x'}$ are

$$p(y'|x') < m_x \leq p(y|x)$$

and

$$p(y'|x) \leq m_{x'} < p(y|x').$$

On the other hand, our upper bound in Equation 15 is informative[9] if and only if[10]

$$p(x, y) + p(x')M_x < p(y)$$

or

$$p(x', y) + p(x)M_{x'} < p(y)$$

which occurs if and only if $p(y|x) < p(y|x')$ or $p(y|x') < p(y|x)$.[11] Therefore, our upper bound is always informative, and thus the informative regions for M_x and $M_{x'}$ coincide with their possible regions.

We illustrate our method for sensitivity analysis of $p(\text{immunity})$ with the following fictitious epidemiological example.

Example 4 *Consider a population consisting of a majority and a minority group. Let the binary random variable U represent the group an individual belongs to. Let X represent whether the individual gets treated or not for a certain disease. Let Y represent whether the individual survives the disease. Assume that the scientific community agrees that U is a confounder for X and Y. Assume also that it is illegal to store the values of U, to avoid discrimination complaints. In other words, the identity of the confounder is known but its values are not. More specifically, consider the following unknown data generation model:*

$$p(u) = 0.2 \qquad p(x|u) = 0.4 \qquad p(y|x, u) = 0.9 \qquad p(y|x, u') = 0.8$$
$$p(x|u') = 0.2 \qquad p(y|x', u) = 0.2 \qquad p(y|x', u') = 0.7.$$

Since this model does not specify the functional forms of the causal mechanisms, we cannot compute the true $p(\textit{immunity})$ (Pearl, 2009). However, we can bound it by Equation 7 and the fact that $p(y_x) = \sum_u p(y|x, u)p(u)$ (Pearl, 2009), which yields $p(\textit{immunity}) \in [0.42, 0.6]$. Note that these bounds cannot be computed in practice because U is unmeasured.

Figure 6 (top) shows the lower bound of $p(\textit{immunity})$ in Equation 15 as a function of the sensitivity parameters m_x and $m_{x'}$. The axes span the possible regions of the parameters. The dashed lines indicate the informative regions of the parameters. Specifically, the bottom left quadrant corresponds to the non-informative region, i.e. the lower bound is zero. In the data generation model considered, $m_x = 0.8$ and $m_{x'} = 0.2$. These values are unknown to the epidemiologist, because U is unobserved. However, the figure reveals that the epidemiologist only needs to have some rough idea of these values to confidently conclude that $p(\textit{immunity})$

9. Note that we already know that $p(\text{immunity}) \leq p(y)$ by Equation 7.
10. Note that the third row in the minimum equals the first plus the second minus the fourth rows.
11. To see it, rewrite $p(y) = p(x, y) + p(x', y)$ and recall Equation 16.

is lower bounded by 0.2. Figure 6 (bottom) shows our upper bound of $p(\textit{immunity})$ *in Equation 15 as a function of the sensitivity parameters* M_x *and* $M_{x'}$*. Likewise, having some rough idea of the unknown values* $M_x = 0.9$ *and* $M_{x'} = 0.7$ *enables the epidemiologist to confidently conclude that the* $p(\textit{immunity})$ *is upper bounded by 0.65. Applying Equation 7 with just observational data produces looser bounds, namely 0 and 0.67. Recall that* $p(\textit{immunity}) \in [0.42, 0.6]$ *in truth.*

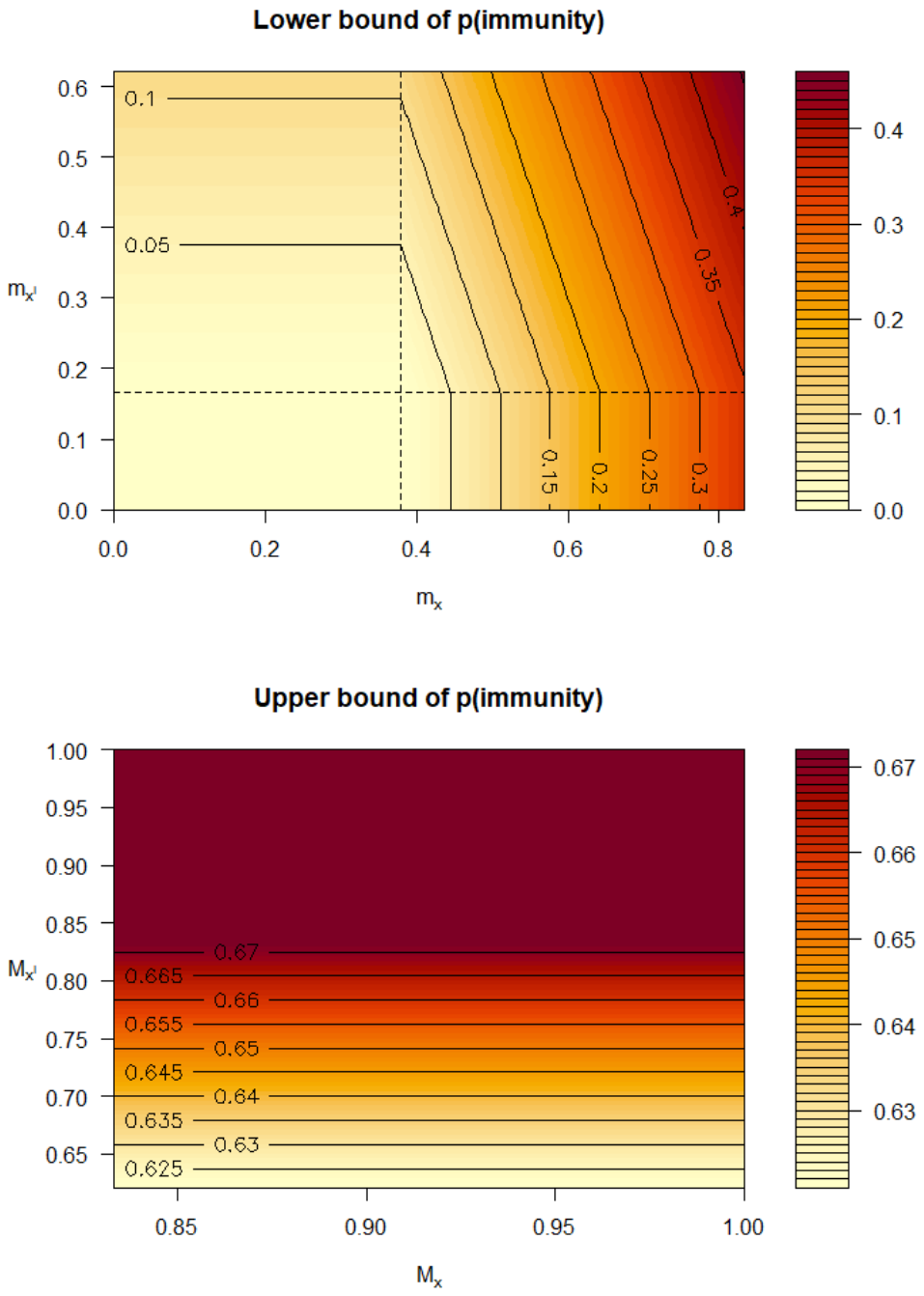


Figure 6: Lower and upper bounds of p(immunity) as functions of the sensitivity parameters m_x, $m_{x'}$, M_x and $M_{x'}$.