Estimating Causal Effects by Bounding Confounding

Philipp Geiger, Dominik Janzing, Bernhard Schölkopf

Max Planck Institute for Intelligent Systems, Tübingen, Germany
{pgeiger, janzing, bs}@tuebingen.mpg.de

Abstract

Assessing the causal effect of a treatment variable X on an outcome variable Y is usually difficult due to the existence of unobserved common causes. Without further assumptions, observed dependences do not even prove the existence of a causal effect from X to Y. It is intuitively clear that strong statistical dependences between X and Y do provide evidence for X influencing Y if the influence of common causes is known to be weak. We propose a framework that formalizes effect versus confounding in various ways and derive upper/lower bounds on the effect in terms of a priori given bounds on confounding. The formalization includes information theoretic quantities like information flow and causal strength, as well as other common notions like effect of treatment on the treated (ETT). We discuss several scenarios where upper bounds on the strength of confounding can be derived. This justifies to some extent human intuition which assumes the presence of causal effect when strong (e.g. close to deterministic) statistical relations are observed.

1 INTRODUCTION

In many situations one wants to estimate the causal effect from an observable X to an observable Y, e.g. if/to what extent smoking causes lung cancer. It is widely agreed that randomized experiments constitute the gold standard for inferring the causal effect. The reason for this is that an ideal randomized experiment excludes the possibility of a (partially) unobserved confounding cause U. However, in many cases conducting randomized experiments would be very expensive or impossible. In these cases, if we do not have any additional knowledge on the setting, then inference of the (precise or approximate) causal effect is generally deemed impossible. In case we have additional knowledge however, it may be possible to *estimate* the causal effect, i.e. derive (upper and/or lower) *bounds* for it. For instance it is well known that if we observe an instrumental variable Z together with X and Y, those bounds can be derived (see e.g. [Pearl, 2000]).

1.1 THE FORMAL FRAMEWORK

To make our discussion as precise as possible we will from this point on use the framework for causality developed in [Pearl, 2000]. Particularly we will make use of the do-calculus formalizing interventions on variables. (It should be mentioned though that we slightly deviate from Pearl's definition of the do-operator as we will further explicate in Section 2.1.)

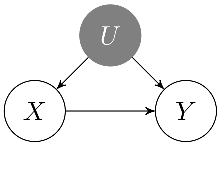

Figure 1: Causal DAG for the confounding scenario (gray means unobserved).

The causal DAG in Figure 1 formalizes the causal structure underlying U, X, Y. Note that we allow U and in some cases also X, Y to be multivariate. Furthermore keep in mind that in some scenarios discussed in the paper we assume U to be partially observed. Our general goal is to estimate the causal effect from X to Y. Formally, this means that we want to estimate $P(Y|\mathrm{do}\,X{=}x)$ or related quantities such as the effect of treatment on the treated (ETT) [Pearl, 2000] or the causal strength from X to Y [Janzing et al., 2013]. Without further assumptions, these quantities are impossible to estimate. To give an extreme example, one can imagine observing the deterministic relationship $P(y|x) = \delta_{yx}$, with δ_{yx} denoting the Kronecker delta. This observation can be induced by two completely different underlying causal structures, the first one being that Y in fact is produced by copying X, the second one being that both X and Y are copied from U without X having any causal effect on Y.

1.2 RELATED WORK

Several approaches have been developed to identify or estimate causal effects in spite of hidden confounders.

Back-door/front-door criterion (see [Pearl, 2000, 2009]): This approach applies for the case where some variables on the confounding path or between X and Y are measured and we know the causal structure underlying all variables together. There are several criteria that allow to decide whether the causal effect from X to Y is identifiable. Furthermore, formulas are available to calculate the effect in these cases. Besides the natural limitation namely requiring a lot of information on additional variables and structures, one drawback of this method is that it cannot be used if X is deterministically coupled to the back-door variable.

Instrumental variables (see e.g. [Pearl, 2000, Angrist et al., 1996, Efron and Feldman, 1991]): In the simplest case, the causal DAG in Figure 1 is augmented by a parentless node Z with an arrow to X. An important example are clinical trials with partial compliance. The additional Z allows to infer bounds on the average causal effect. One drawback of this method is that it yields a convex optimization problem with the number of equations growing exponentially with the cardinality of X. Furthermore, to apply this method one needs to know $p(X, Y|Z)$ while in Section 4.2 we present a scenario where $p(Z)$ (additional to $p(X, Y)$) helps to estimate the causal effect.

Regression discontinuity design (see e.g. [Thistlewaite and Campbell, 1960, Imbens and Lemieux, 2008, Lee and Lemieux, 2010]): This framework is applicable to cases where an additional observable Z mediating between U and X is measured and X is a deterministic function of Z that contains a discontinuity. Under the assumption of linearity of the remaining structural equations, the effect from X to Y, i.e. the linear coefficient, can be identified. One limitation of this method is that it needs the discontinuity and a high slope alone does not suffice.

1.3 OUR APPROACH

The approach we suggest to estimate causal effects in spite of confounding consists of two parts: In the *first part* (Section 3) we propose various possibilities to formalize the following notions:

- *Observed dependence:* the dependence of Y on X that we can observe based on $p(X, Y)$.
- *Back-door dependence:* the "spurious association" [Pearl, 2000] between X and Y due to the confounder U.
- *Causal effect:* what happens to Y upon intervening on X; this includes notions of conditional causal effect such as the ETT.

For all formalizations we present inequalities (see Table 1 for an overview) which turn out to always have the following prototypical form:

$$\begin{bmatrix}\text{back-door}\\ \text{dependence}\end{bmatrix} \geq \mathrm{d}\left(\begin{bmatrix}\text{observed}\\ \text{dependence}\end{bmatrix}, \begin{bmatrix}\text{causal}\\ \text{effect}\end{bmatrix}\right)$$

(where the $\mathrm{d}(\cdot,\cdot)$ stands for deviation measure). In some of these results, observed dependence, back-door dependences, and causal effect are real numbers and $\mathrm{d}(\cdot,\cdot)$ simply stands for the usual difference which allows us to convert the prototypical form into

$$\begin{bmatrix}\text{causal}\\ \text{effect}\end{bmatrix} \geq \begin{bmatrix}\text{observed}\\ \text{dependence}\end{bmatrix} - \begin{bmatrix}\text{back-door}\\ \text{dependence}\end{bmatrix},$$

which may be more convenient for applications.

In order to draw conclusions on the true causal effect using the inequalities from the first part, one needs to have knowledge on the back-door dependence. Therefore, in the *second part* (Section 4), we demonstrate how in various situations one can come up with bounds on the back-door dependence. Based on these together with the observed dependence one can then infer bounds on the true causal effect.

Before getting started, in Section 2, we present several definitions which are needed throughout the paper.

2 PREREQUISITES

Keep in mind the following definitions and results throughout the paper.

2.1 DEFINITION OF CAUSAL MODEL AND do-OPERATOR

As already mentioned we essentially use the framework developed in [Pearl, 2000] to discuss causal relations. Let $V = \{X_1, \ldots, X_n\}$ be a set of random variables. A *causal model* M w.r.t. V consists of a DAG G, noise variables N_i for each i with a joint distribution P that makes them jointly independent, and structural equations $X_i := f_i(\mathrm{PA}_i^G, N_i)$ for each i, where PA_i^G denotes the parents of X_i in G.

Table 1: Formalizing observed dependence (O), back-door dependence (B), causal effect (C) and deviation measure (d). $\mathfrak{C}_{X \to Y}$ denotes the strength of the influence of X on Y in the sense of [Janzing et al., 2013], $I(X \to Y | \mathrm{do}\, U)$ is the information flow by [Ay and Polani, 2008], $\mathrm{D}[\cdot \| \cdot]$ denotes Kullback-Leibler divergence [Cover and Thomas, 1991]. The symbol do is the do-operator defined by [Pearl, 2000].

Sec.	**O/B/C/d formalized by ...**	
3.1	O	$\mathrm{I}(X : Y)$
	B	$\mathrm{I}(U : X)$, $\mathfrak{C}_{U \to X}$
	C	$\mathfrak{C}_{X \to Y}$
	d	difference
3.2	O	$\mathrm{I}(X : Y)$
	B	$\mathrm{I}(U : X)$
	C	$\mathrm{I}(X \to Y \mid \mathrm{do}\, U)$
	d	difference
3.3	O	$p(Y \mid X{=}x)$
	B	$\mathrm{I}(X : U)$, $\min\{\mathfrak{C}_{U \to X}, \mathfrak{C}_{U \to Y}\}$
	C	$p(Y \mid \mathrm{do}\, X{=}x)$
	d	$\mathrm{D}[\cdot \| \cdot]$
3.4	O	$\mathbb{E}[\mathrm{d}_x \log p(Y \mid X{=}x)^2]$
	B	$\mathbb{E}[\partial_2 \log p(Y \mid X{=}x, \mathrm{do}\, X{=}x))^2]$
	C	$\mathbb{E}[\partial_1 \log p(Y \mid X{=}x, \mathrm{do}\, X{=}x))^2]$
	d	difference
3.5	O	$\mathbb{E}[Y \mid X{=}x'] - \mathbb{E}[Y \mid X{=}x]$
	B	$\mathbb{E}[Y \mid X{=}x', \mathrm{do}\, X{=}x]$
		$-\mathbb{E}[Y \mid X{=}x, \mathrm{do}\, X{=}x]$
	C	$\mathbb{E}[Y \mid X{=}x', \mathrm{do}\, X{=}x']$
		$-\mathbb{E}[Y \mid X{=}x', \mathrm{do}\, X{=}x]$
	d	difference
3.6	O	$\mathrm{d}_x \mathbb{E}[Y \mid X{=}x]$
	B	$\partial_1 \mathbb{E}[Y \mid X = x, \mathrm{do}\, X{=}x]$
	C	$\partial_2 \mathbb{E}[Y \mid X = x, \mathrm{do}\, X{=}x]$
	d	difference

Now we define the do-*operator*. Given any $X_i \in V$ the post-intervention causal model $M_{\mathrm{do}\, X_i = x'}$ is obtained from M in the following way: For each child X_j of X_i, we replace the structural equation $X_j = f_j(\mathrm{PA}_j^G \backslash X_i, X_i, N_j)$ by $X_j = f_j(\mathrm{PA}_j^G \backslash X_i, x', N_j)$. Note that the random variable X_i stays in the model it just no longer has children. (This is the point where we deviate from [Pearl, 2000]. Note the analogy to the splitting of nodes in [Richardson and Robins, 2013, Robins et al., 2007], where X_i is replaced with two deterministically coupled variables, one being adjacent to the parents of X_i and one to the children of X_i. Then we refer to an intervention on the latter one while the first one is kept.)

The new set of structural equations of $M_{\mathrm{do}\, X_i = x'}$ together with the noise variables N_i for all i induce a new joint distribution on $X_1, \ldots, X_n$ which we denote by $P(X_1, \ldots, X_n | \mathrm{do}\, X{=}x')$. In particular, given M contains the variables X and Y, quantities such as $P(Y | X{=}x, \mathrm{do}\, X{=}x')$ are well defined. (Note that $P(Y | X{=}x, \mathrm{do}\, X{=}x')$, based on our definition of $M_{\mathrm{do}\, X = x'}$, coincides with the counterfactual distribution $P(Y_{x'} | X{=}x)$ as defined in [Pearl, 2000].)

2.2 DISTRIBUTIONS AND DENSITIES

Throughout the paper we will work with U, X, Y with discrete as well as with continuous ranges.

Unless noted otherwise we make the following fundamental assumption regarding the distributions of the random variables in a causal model M with causal DAG G: for each $X_j \in V$, the random variable $f_j(\mathrm{pa}_j, N_j)$ has a density w.r.t. the Lebesgue measure (in the continuous case) or w.r.t. the counting measure (in the discrete case) respectively, denoted by $q_j(x_j; \mathrm{pa}_j)$ *for each value* pa_j *of* PA_j (note that we have to slightly deviate from this assumption in Section 4.1 though). This assumption implies the following simple lemma, which is only formulated for the case $n = 3$, since we only need this case in the present paper. A proof can be found in the supplement.

Lemma 1. *Under the assumption made above, the joint distribution of X_1, X_2, X_3 induced by a causal model M or any post-interventional model $M_{\mathrm{do}\, X_i = x}$ has a density w.r.t. the Lebesgue measure (in the continuous case) or counting measure (in the discrete case), respectively. Moreover, this density factorizes according to the causal DAG belonging to the respective model.*

Regarding any causal model M with causal DAG as depicted in Figure 1, also the following simple statement holds true. The proof is obvious but we present it in the supplement anyway. Note that the lemma is similar to [Pearl, 2000, Corollary 7.3.2], but better tailored for our definition of $M_{\mathrm{do}\, X = x}$.

Lemma 2. *For all x we have*

$$p(Y | X = x, \mathrm{do}\, X{=}x) = p(Y | X = x),$$
$$\mathbb{E}[Y | X = x, \mathrm{do}\, X{=}x] = \mathbb{E}[Y | X = x].$$

3 THE RELATION BETWEEN OBSERVED DEPENDENCE, BACK-DOOR DEPENDENCE AND CAUSAL EFFECT

In this section we present various possibilities to formalize the notions of observed dependence, back-door dependence and causal effect. For all formalizations we prove that the back-door dependence is equal to or upper bounds the deviation between the observed dependence and the actual causal effect.

Subsections 3.1, 3.2 apply to X, Y, U with finite range. Subsections 3.3, 3.5 apply to X, Y, U with arbitrary range. Subsections 3.4 and 3.6 apply to X with continuous range.

Keep in mind that $\mathrm{H}(\cdot)$ denotes the Shannon entropy, $\mathrm{I}(\cdot:\cdot)$ $(\mathrm{I}(\cdot:\cdot|\cdot))$ the (conditional) mutual information, and $\mathrm{D}[\cdot\|\cdot]$ the Kullback-Leibler divergence, all based on logarithms with base 2. For details see [Cover and Thomas, 1991].

3.1 ESTIMATING THE CAUSAL STRENGTH FROM X TO Y

The basic quantities in this section are:
- observed dep.: $\mathrm{I}(X:Y)$,
- back-door dep.: $\mathrm{I}(X:U)$, $\mathfrak{C}_{U\to X}$,
- causal effect: $\mathfrak{C}_{X\to Y}$.

We consider the case of U, X, Y having finite range. [Janzing et al., 2013] proposed a definition for the causal strength of a set of arrows in a causal DAG.

We briefly want to repeat this definition for the special case of measuring the strength of a single arrow. For a set of observables $V = \{X_1, \dots, X_n\}$, a DAG G' with V as the set of nodes and a joint distribution $p(X_1, \dots, X_n)$ and for any arrow $X_i \to X_j$ in G' we first define the distribution $p_{X_i\to X_j}$ corresponding to deleting $X_i \to X_j$ from the graph and feeding X_j with an independent copy of X_i instead, see also [Ay and Krakauer, 2007]:

$$\begin{aligned}
&p_{X_i\to X_j}(x_j|\mathrm{pa}_{X_j}^{X_i\to X_j}) := \sum_{x_i'} p(x_i')p(y|x_i', \mathrm{pa}_{X_j}^{X_i\to X_j}),\\
&p_{X_i\to X_j}(x_k|\mathrm{pa}_{X_k}^{X_i\to X_j}) := p(x_k|\mathrm{pa}_{X_k}), \text{ for all } k \neq j,\\
&p_{X_i\to X_j}(x_1, \dots, x_n) := \prod_{k=1}^{n} p_{X_i\to X_j}(x_k|\mathrm{pa}_{X_k}^{X_i\to X_j}),
\end{aligned}$$

where $\mathrm{pa}_{X_k}^{X_i\to X_j}$ denotes (values of) the set of parents of X_k in the modified graph G' without arrow $X_i \to X_j$ (obviously this actually only makes a change for pa_{X_j}). Now we are able to define the *causal strength* $\mathfrak{C}_{X_i\to X_j}$ by the impact of the edge deletion:

$$\mathfrak{C}_{X_i\to X_j} := \mathrm{D}[p(X_1, \dots, X_n)\|p_{X_i\to X_j}(X_1, \dots, X_n)].$$

Let us get back to our specific confounding scenario (the causal DAG in Figure 1). For general DAGs, [Janzing et al., 2013] shows $\mathfrak{C}_{X\to Y} \geq I(X:Y|PA_Y \setminus X)$, that is, the information Y contains about X given its other parents is a lower bound for causal strength (they argue that this property would be desirable for other information-theoretic measures of causal strength as well). Hence in our confounding scenario (Figure 1) we have $\mathfrak{C}_{X\to Y} \geq \mathrm{I}(X:Y|U)$. Also keep in mind that $\mathfrak{C}_{U\to X} = \mathrm{I}(U:X)$ in our setting.

Lemma 3. *We have*

$$\mathrm{I}(X:Y|U) \geq \mathrm{I}(X:Y) - \mathrm{I}(X:U). \tag{1}$$

Proof. The statement follows from the fact that $\mathrm{I}(X:Y|U) + \mathrm{I}(X:U) = \mathrm{I}(X:U,Y) \geq \mathrm{I}(X:Y)$. □

We consider $\mathrm{I}(X:Y)$ as a measure of *observed dependence* between X and Y. The following theorem shows that the *back-door dependence* $\mathfrak{C}_{U\to X}$ bounds the difference between the observed dependence and the true *causal effect* $\mathfrak{C}_{X\to Y}$.

Theorem 1. *We have*

$$\mathfrak{C}_{U\to X} \geq \mathrm{I}(X:Y) - \mathfrak{C}_{X\to Y}. \tag{2}$$

Proof. This follows from Lemma 3 together with the fact that $\mathfrak{C}_{X\to Y} \geq \mathrm{I}(X:Y|U)$ and $\mathfrak{C}_{U\to X} = \mathrm{I}(X:U)$ in our confounding scenario (i.e. the DAG in Figure 1). □

3.2 ESTIMATING THE INFORMATION FLOW FROM X TO Y

The basic quantities in this section are:
- observed dep.: $\mathrm{I}(X:Y)$,
- back-door dep.: $\mathrm{I}(X:U)$,
- causal effect: $\mathrm{I}(X\to Y|\mathrm{do}\,U)$.

Another information theoretic quantity to measure the causal effect of X on Y is the *information flow* proposed by [Ay and Polani, 2008]. In our setting (the causal DAG in Figure 1) it is defined as

$$\begin{aligned}
&\mathrm{I}(X\to Y|\mathrm{do}\,U) :=\\
&\sum_u p(u)\sum_x p(x|\mathrm{do}\,U{=}u)\sum_y p(y|\mathrm{do}\,X{=}x, \mathrm{do}\,U{=}u)\\
&\times \log\frac{p(y|\mathrm{do}\,X{=}x, \mathrm{do}\,U{=}u)}{\sum_{x'} p(y|\mathrm{do}\,X{=}x', \mathrm{do}\,U{=}u)p(x'|\mathrm{do}\,U{=}u)}.
\end{aligned}$$

Since $p(y|\mathrm{do}\,X{=}x, \mathrm{do}\,U{=}u) = p(y|x,u)$ in our setting, we simply have $\mathrm{I}(X\to Y|\mathrm{do}\,U) = \mathrm{I}(X:Y|U)$.

So we can establish a theorem for the information flow similar to the one for the causal strength. it follows immediately from Lemma 3.

Theorem 2. *We have*

$$\mathrm{I}(X:U) \geq \mathrm{I}(X:Y) - \mathrm{I}(X\to Y|\mathrm{do}\,U). \tag{3}$$

3.3 BOUNDING THE KULLBACK-LEIBLER DIVERGENCE BETWEEN $p(Y|X{=}x)$ AND $p(Y|\mathrm{do}\,X{=}x)$

The basic quantities in this section are:
- observed dep.: $p(Y|X{=}x)$,
- back-door dep.: $\mathrm{I}(X:U)$, $\min\{\mathfrak{C}_{U\to X}, \mathfrak{C}_{U\to Y}\}$,
- causal effect: $p(Y|\mathrm{do}\,X{=}x)$.

In some sense, $p(Y|\mathrm{do}\,X{=}x)$ is the most fundamental characterization of the *causal effect* from X to Y, while $p(Y|X{=}x)$ can be seen as the corresponding characterization of their *observed dependence*. In this section we show that the deviation between these two objects can be bounded by quantities which measure the *back-door dependence*, $\mathrm{I}(X:U)$ and $\min\{\mathfrak{C}_{U\to X}, \mathfrak{C}_{U\to Y}\}$. We formalize the notion of deviation here by

$$\begin{aligned}&\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)]\\ &:= \sum_x p(x)\mathrm{D}[p(Y|x)\|p(Y|\mathrm{do}\,X{=}x)].\end{aligned}$$

Theorem 3. *We have*

$$\begin{aligned}\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)] &\le \min\{\mathfrak{C}_{U\to X}, \mathfrak{C}_{U\to Y}\}\\ &\le \mathrm{I}(X:U).\end{aligned}$$

Proof. First note that
$p_{U\to X}(u,x,y) = p(u)p(x)p(y|u,x)$ and
$p_{U\to Y}(u,x,y) = p(u)p(x|u)\sum_{u'} p(y|u,x)p(u')$.

This implies
$p(y|\mathrm{do}\,X{=}x) = p_{U\to X}(y|X{=}x)$ and
$p(y|\mathrm{do}\,X{=}x) = p_{U\to Y}(y|X{=}x)$.

Therefore, using the chain rule for Kullback-Leibler divergence,

$$\begin{aligned}&\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)] = \mathrm{D}[p(Y|X)\|p_{U\to X}(Y|X)]\\ &\le \mathrm{D}[p(X,Y)\|p_{U\to X}(X,Y)] = \mathfrak{C}_{U\to X} (= \mathrm{I}(U:X)).\end{aligned}$$

Similarly one can derive $\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)] \le \mathfrak{C}_{U\to Y}$. □

The above theorem makes a statement w.r.t. the divergence between $p(Y|x)$ and $p(Y|\mathrm{do}\,X{=}x)$ *averaged* over all values x of X. But it is also possible to derive a pointwise version:

Theorem 4. *For all x*

$$\mathrm{D}[p(Y|x)\|p(Y|\mathrm{do}\,x)] \le \mathrm{D}[p(U|x)\|p(U)],$$

with equality iff $p(u|x) = p(u)$ for all u.

Proof. By the log sum inequality we have

$$\begin{aligned}&p(y|x)\log\frac{p(y|x)}{p(y|\mathrm{do}\,x)}\\ &= \left(\sum_u p(y|x,u)p(u|x)\right)\log\frac{\sum_u p(y|x,u)p(u|x)}{\sum_u p(y|x,u)p(u)}\\ &\le \sum_u p(y|x,u)p(u|x)\log\frac{p(y|x,u)p(u|x)}{p(y|x,u)p(u)} \qquad (4)\\ &= \sum_u p(y,u|x)\log\frac{p(u|x)}{p(u)}.\end{aligned}$$

Equality holds in (4) iff $p(y|x,u)p(u|x) = cp(y|x,u)p(u)$ for all u and some constant c, i.e. iff $p(u|x) = p(u)$ for all u. Summing over all y yields the claimed inequality. □

Note that taking the average w.r.t. X in Theorem 4 is another way to prove the first part of Theorem 3. With a similar proof we can also derive the following inequality w.r.t. the "inverse mutual information" $\mathrm{D}[p(U)p(X)\|p(U,X)]$ (as opposed to the usual mutual information $\mathrm{I}(U:X) = \mathrm{D}[p(U,X)\|p(U)p(X)]$). For this purpose let us define

$$\begin{aligned}&\mathrm{D}[p(Y|\mathrm{do}\,X)\|p(Y|X)]\\ &:= \sum_x p(x)\sum_y p(y|\mathrm{do}\,X{=}x)\log\frac{p(y|\mathrm{do}\,X{=}x)}{p(y|X{=}x)}.\end{aligned}$$

Corollary 1. *We have*

$$\mathrm{D}[p(Y|\mathrm{do}\,X)\|p(Y|X)] \le \mathrm{D}[p(U)p(X)\|p(U,X)].$$

To assess which bound is relevant for a scenario, we recall that for two distributions p and q, $\mathrm{D}[p\|q]$ diverges when $q = 0$ and $p > 0$ on a set of Lebesgue measure greater than 0. If the observed dependence $p(Y|X)$ is deterministic, $p(Y|\mathrm{do}\,X)$ needs to be deterministic if $\mathrm{D}[p(Y|\mathrm{do}\,X)\|p(Y|X)]$ is finite.

3.3.1 An example for bounding the average causal effect from X to Y

Often one is interested in estimating the *average causal effect* $\mathbb{E}[Y|\mathrm{do}\,X{=}x'] - \mathbb{E}[Y|\mathrm{do}\,X{=}x]$ for two values x, x' of X [Pearl, 2000], in particular because this quantity is easy to interpret. In what follows, we want to give an example how one can derive bounds on this quantity based on Theorem 3. It is important to mention however, that the assumptions we make are very restrictive. The purpose of the example is only to show that information theoretic bounds on the back-door dependence *can*, under appropriate assumptions, imply bounds for the average causal effect.

Let X be binary, $p(Y|x) = \mathcal{N}(\mu_x, \sigma^2)$, and $p(Y|\mathrm{do}\,X{=}x) = \mathcal{N}(\mu_{\mathrm{do}\,x}, \sigma^2_{\mathrm{do}})$, for $x = 0, 1$ (hence particularly $\mathbb{E}[Y|\mathrm{do}\,X{=}x] = \mu_{\mathrm{do}\,x}$).[1]

In this case we can calculate (have in mind that ln is the natural logarithm)

$$\begin{aligned} &p(X{=}0)(\mu_0 - \mu_{\mathrm{do}\,0})^2 + p(X{=}1)(\mu_1 - \mu_{\mathrm{do}\,1})^2 \\ &= 2\sigma^2_{\mathrm{do}}\left(\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)] - \ln\frac{\sigma^2_{\mathrm{do}}}{\sigma^2} - \frac{\sigma^2}{2\sigma^2_{\mathrm{do}}} + \frac{1}{2}\right) \\ &\leq 2\sigma^2_{\mathrm{do}}\left(\min\{\mathfrak{C}_{U\to X}, \mathfrak{C}_{U\to Y}\} - \ln\frac{\sigma^2_{\mathrm{do}}}{\sigma^2} - \frac{\sigma^2}{2\sigma^2_{\mathrm{do}}} + \frac{1}{2}\right). \end{aligned} \tag{5}$$

Now assume we fix $\min\{\mathfrak{C}_{U\to X}, \mathfrak{C}_{U\to Y}\}$ and σ^2_{do}. Keep in mind that μ_0, μ_1, σ^2 are observed. Then we can derive upper and lower bounds on the average causal effect $\mu_{\mathrm{do}\,1} - \mu_{\mathrm{do}\,0}$ by maximizing and minimizing it, respectively, under the constraints on the pair $(\mu_{\mathrm{do}\,1}, \mu_{\mathrm{do}\,0})$ imposed by inequality (5).

3.4 ESTIMATING THE FISHER INFORMATION

The basic quantities in this section are:
- observed dep.: $\mathcal{F}_{Y|X}(x)$,
- back-door dep.: $\mathcal{F}^1_{Y|X,\mathrm{do}\,X}(x, x)$,
- causal effect: $\mathcal{F}^2_{Y|X,\mathrm{do}\,X}(x, x)$.

In the following, $\partial_i f(x, x')$, $i = 1, 2$, denotes the partial derivative w.r.t. the ith argument of f evaluated at position (x, x'). And $\mathrm{d}_x g(x)$ denotes the total derivative of $g(x)$ w.r.t. x at position x, in particular $\mathrm{d}_x f(x, x) = \mathrm{d}_x g(x)$ for $g(x) := f(x, x)$.

Given a family of distributions depending on continuous parameters, *Fisher information* provides a natural way to quantify the sensitivity of a probability distribution to infinitesimal parameter changes. It plays an important role for the error made when estimating the true parameter value from empirical data [Rao, 1945]. Here we quantify causal influence by the sensitivity of $p(Y|\mathrm{do}\,x)$ to small changes of x. This can be considered as a "local" measure of causal strength in the neighborhood of x. We introduce the following notation:

$$\mathcal{F}_{Y|X}(x) := \int (\mathrm{d}_x \log p(y|X{=}x))^2 p(y|X{=}x)\mathrm{d}y,$$

$$\mathcal{F}^i_{Y|X,\mathrm{do}\,X}(x, x') :=$$

$$\int (\partial_i \log p(y|X{=}x, \mathrm{do}\,X{=}x'))^2 p(y|X{=}x, \mathrm{do}\,X{=}x')\mathrm{d}y,$$

for $i = 1, 2$.

Theorem 5. *For all x*

$$\sqrt{\mathcal{F}_{Y|X}(x)} - \sqrt{\mathcal{F}^2_{Y|X,\mathrm{do}\,X}(x, x)} \leq \sqrt{\mathcal{F}^1_{Y|X,\mathrm{do}\,X}(x, x)}.$$

A proof can be found in the supplement.

3.5 ESTIMATING THE EFFECT OF TREATMENT ON THE TREATED FROM X TO Y

The basic quantities in this section are:
- observed dep.:
$\mathbb{E}[Y|X{=}x'] - \mathbb{E}[Y|X{=}x]$,
- back-door dep.:
$\mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x] - \mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]$,
- causal effect:
$\mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x'] - \mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x]$.

Following [Pearl, 2000], we define the quantity

$$\mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x'] - \mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x]$$

as the *effect of treatment on the treated*. As the name already suggests, the idea behind this quantity is to measure the deviation between the average response of the treated subjects and the average response of these same subjects had they not been treated. The following result w.r.t. the effect of treatment on the treated follows from Lemma 2.

Theorem 6. *We have for all x, x'*

$$\begin{aligned} &\mathbb{E}[Y|X{=}x'] - \mathbb{E}[Y|X{=}x] \\ &= \mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x'] - \mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x] \\ &\quad + \mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x] - \mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]. \end{aligned}$$

Note that in mediation analysis [Pearl, 2001, Avin et al., 2005, Robins and Greenland, 1992] a similar splitting into direct and indirect effect is used. However mediation analysis addresses the problem of defining direct and indirect *causal* effects and not back-door dependences.

We briefly want to discuss the other quantities that appear in the theorem. Obviously, $\mathbb{E}[Y|X{=}x'] - \mathbb{E}[Y|X{=}x]$ measures the *observed dependence* of Y on X. Now keep in mind that in $M_{\mathrm{do}\,X=x}$, X has no causal effect on Y anymore and hence Y statistically depends on X solely via U. Therefore the difference

$$\mathbb{E}[Y|X{=}x', \mathrm{do}\,X{=}x] - \mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]$$

measures the strength of the *back-door dependence* of Y on X.

[1] Note, however, that both $p(Y|X{=}0)$ and $p(Y|X{=}1)$ being Gaussian actually provides some evidence for the absence of confounding since a confounder will often destroy this simple structure of $P(Y|X)$ [Janzing et al., 2011].

3.6 ESTIMATING THE DIFFERENTIAL EFFECT OF TREATMENT ON THE TREATED FROM X TO Y

The basic quantities in this section are:
- observed dep.: $\mathrm{d}_x\mathbb{E}[Y|X{=}x]$,
- back-door dep.: $\partial_1\mathbb{E}[Y|X = x, \mathrm{do}\,X{=}x]$,
- causal effect: $\partial_2\mathbb{E}[Y|X = x, \mathrm{do}\,X{=}x]$.

First note that by $\partial_i\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x']$ we mean $\partial_i f(x, x')$ for $f(x, x') := \mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x']$ (recall that ∂_i denotes the partial derivative w.r.t. the ith argument). In the case of continuous random variables U, X, Y we want to consider the following quantity (if it exists i.e. if the conditional expectation is differentiable)

$$\partial_2\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x],$$

which we call *differential effect of treatment on the treated* or simply *differential effect* in cases where this does not lead to confusions. It is the analog to the discrete effect of treatment on the treated (see Section 3.5) for the case of infinitesimal interventional changes on X; we simply replaced a difference by a derivative.

Similar to Theorem 6 we can establish the following result. It follows from the chain rule for derivatives together with Lemma 2.

Theorem 7. *For all x*

$$\begin{aligned}\mathrm{d}_x\mathbb{E}[Y|X{=}x] = \;&\partial_1\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]\\ &+ \partial_2\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x].\end{aligned}$$

The interpretation of this theorem is similar to the one for Theorem 6. Obviously, $\mathrm{d}_x\mathbb{E}[Y|X{=}x]$ is the *observed dependence*, whereas the quantity $\partial_1\mathbb{E}[Y|X = x, \mathrm{do}\,X{=}x]$ measures the *back-door dependence* of Y on X. So the observed dependence of Y on X splits into the causal effect plus the back-door dependence.

4 PROTOTYPICAL SCENARIOS WITH BOUNDS ON THE BACK-DOOR DEPENDENCE

In this section we present several prototypical scenarios where bounds on the back-door dependence between X and Y can be derived. Together with our results from Section 3 these bounds help to estimate causal effect from X to Y.

4.1 A QUALITATIVE TOY EXAMPLE

We want to give an example that demonstrates how human intuition concerning observed dependence and causal effect relates to the theorems from Section 3. Assume there is a drug that is indicated for a specific disease. We observe some not too small number of people with the disease and see that some of them take the drug and some do not. We find that all persons who took the drug recovered on the same day whereas none of the persons not taking the drug recovered that fast. For each sick person let X denote the date he or she takes the drug and Y the date he or she recovers. Since these are only observations, we cannot exclude that there is a confounder U, i.e. we assume the usual causal DAG (Figure 1). We estimate the distribution of Y given X by the empirical distribution, i.e. $p(y|x) = \delta_{yx}$, where δ_{yx} denotes the Kronecker delta.

Given the above setting probably most people would argue that there has to be some effect from the drug to the immediate healing of those people who took it. However, formally and without further assumptions $p(Y|x)$ alone does not even tell us if there is a causal link from X to Y at all. With the help of Theorem 3 though, we can formally reason as follows. We make the weak additional assumption that X cannot be completely determined by U which we formalize by $\mathrm{I}(U : X) < \mathrm{H}(X)$. It seems implausible that there exists a common cause of X and Y that determines both, the exact date X a person takes the drug and the recovering date Y. E.g. the wealth of a person may strongly influence both, the treatment he or she takes and how quickly he or she recovers (via the general health condition), however it seems not plausible that the wealth determines the exact day of taking the drug and of recovering.

For a proof by contradiction we may assume that there is no causal effect from X to Y, i.e. $p(Y|\mathrm{do}\,X{=}x) = p(Y|\mathrm{do}\,X{=}x')$ for all x, x'. Then

$$\begin{aligned}&\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X{=}x)]\\ &= \sum_x p(x)\mathrm{D}[p(Y|X{=}x)\|p(Y|\mathrm{do}\,X{=}x)]\\ &= \sum_x p(x)\sum_y \delta_{yx}\log\frac{\delta_{yx}}{P(Y = y|\mathrm{do}\,X{=}x)}\\ &= \sum_x p(x)\log\frac{1}{p(Y{=}x|\mathrm{do}\,X{=}0)} \geq \mathrm{H}(X),\end{aligned}$$

where the last inequality is due to Gibb's inequality [Cover and Thomas, 1991].

On the other hand, due to Theorem 3 we have

$$\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\,X)] \leq \mathrm{I}(X : U) < \mathrm{H}(X),$$

which yields the contradiction. Hence we could formally show that there has to be some causal effect from X to Y, $p(Y|\mathrm{do}\,X{=}x) \neq p(Y|\mathrm{do}\,X{=}x')$ for some x, x'. Note that the above argumentation completely transfers to any other situation where $p(y|x) = \delta_{yx}$, particularly any other range of X and Y.

4.2 PARTIAL RANDOMIZATION SCENARIO

We first discuss a formal scenario, then an application example, and afterwards we discuss how the scenario and our result is related to the instrumental variable design [Pearl, 2000].

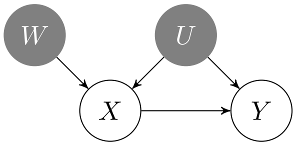

Figure 2: The partial randomization causal DAG.

4.2.1 THE FORMAL PROTOTYPE

We consider a scenario where we have measured X and Y, and where hidden variables U and W are present and we know the distribution of W. The underlying causal structure of all variables looks like the causal DAG depicted in Figure 2. We assume that W is binary. Furthermore we assume that in this scenario $\mathrm{I}(U : X|W = 0) = 0$. The intuition behind this assumption is that W decides whether X is influenced by U ($W = 1$) or not. This scenario implies the following inequality. A proof can be found in the supplement.

Proposition 1. *In the given scenario we have* $\mathrm{I}(U : X) \leq \mathrm{H}(X)p(W{=}1)$.

Now we can employ our results from Sections 3.1 through 3.3. We obtain the following bounds:

$$\mathrm{I}(X : Y) - \mathfrak{C}_{X\to Y} \leq \mathrm{H}(X)p(W{=}1), \quad (6)$$

$$\mathrm{I}(X : Y) - \mathrm{I}(X \to Y|\mathrm{do}\, U) \leq \mathrm{H}(X)p(W{=}1), \quad (7)$$

$$\mathrm{D}[p(Y|X)\|p(Y|\mathrm{do}\, X)] \leq \mathrm{H}(X)p(W{=}1). \quad (8)$$

Note that under strong assumptions, one can also apply the result from Section 3.3.1 to estimate the average causal effect $\mathbb{E}[Y|\mathrm{do}\, X{=}x'] - \mathbb{E}[Y|\mathrm{do}\, X{=}x]$ for two values x, x' of X.

4.2.2 ADVERTISEMENT LETTER EXAMPLE

Assume we are managers of a mail order company, and want to know the effect of sending advertisement letters on the ordering behavior of the recipients. We have a data set of (X, Y) pairs with X denoting whether a letter was sent to a specific person and let Y denote the total costs of the products ordered by this person afterwards (within some fixed time span). Assume we have enough data to estimate $p(X, Y)$. Furthermore, assume that so far there were already imperfect guidelines based on rough intuition on whom to send letters and whom not. These guidelines introduce a potential confounder U since letters were more likely send to potential customers with properties that made them also more likely to order something (if the guidelines were not complete nonsense). It is known however that only some employees sticked to these guidelines. Let W denote whether a letter was sent out in compliance with these guidelines ($W = 1$) or not.

Based on an estimate of how many employees complied with the guidelines, we also have an estimate of $p(W = 1)$, i.e. the fraction of letters that was sent out in compliance with the guidelines. Based on Proposition 1, we know that $\mathrm{I}(U{:}X) \leq \mathrm{H}(X)p(W{=}1)$. Hence we have an upper bound on the back-door dependence of Y on X. Particularly we can apply inequalities (6) to (8) and, under strong additional assumptions, the result w.r.t. the average causal effect from Section 3.3.1.

For example by (6) we have $\mathrm{I}(X{:}Y) - \mathrm{H}(X)p(W{=}1) \leq \mathfrak{C}_{X\to Y}$. Now assume $\mathrm{H}(X){\approx}1$ (we sent a letter to roughly every second person in our register) and $p(W{=}1){\approx}0.5$ (only half the employees sticked to the guidelines). Then if we observe a strong dependence of Y on X, say $\mathrm{I}(X{:}Y){\approx}0.75$, then we can conclude that $\mathfrak{C}_{X\to Y}{\gtrsim}0.25$, i.e. our advertisement letters have a significant effect on the potential customers.

4.2.3 DIFFERENCE TO INSTRUMENTAL VARIABLE DESIGN

We already mentioned the instrumental variable design [Pearl, 2000] in Section 1. In this design it is assumed that an additional variable W is observed such that the causal structure of all variables together is as depicted in Figure 2, except that W is not hidden. The prototypical application scenario for this design are clinical trials with partial compliance. [Pearl, 2000] describes a method to derive bounds on the average causal effect $\mathbb{E}[Y|\mathrm{do}\, X{=}1] - \mathbb{E}[Y|\mathrm{do}\, X{=}0]$. This analysis heavily depends on the range of X, Y, and W and involves convex optimization in 15-dimensional space already for the case where all variables are binary (since U can be assumed to attain 16 different values).

The advantage of our approach lies in the fact that the ranges of the variables may be arbitrary without increasing the complexity – for the cost of getting less tight bounds than an explicit modeling, of course. One can get bounds for the case where neither X nor

Y are binary, e.g., in a drug testing scenario with different doses and descriptions of health conditions that are more complex than just reporting recovery or not. Moreover, we do not need complete knowledge of $p(Y, X|W)$ provided that we have some knowledge on W that provides upper bounds on $\mathrm{I}(X{:}U)$.

4.3 A VARIANT OF THE REGRESSION DISCONTINUITY DESIGN

We already mentioned the regression discontinuity design (RDD) [Thistlewaite and Campbell, 1960, Imbens and Lemieux, 2008, Lee and Lemieux, 2010] in Section 1. It is a quasi-experimental design that can help to estimate the causal effect from X to Y in cases where an additional variable Z is measured and the underlying causal DAG of all variables together is as depicted in Figure 3. The design usually requires that X is a deterministic function of Z that contains a discontinuity, that all remaining structural equations are linear, and that $\mathbb{E}[U|Z = z]$ is continuous in z. (Note that the causal DAG in Figure 3 is a special case of the general confounding scenario depicted in Figure 1, which can be seen by replacing U in Figure 1 by $U' := (U, Z)$.)

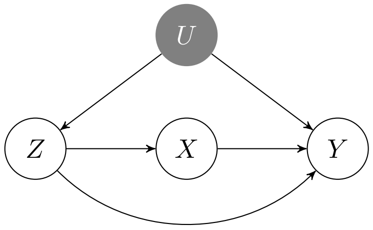

Figure 3: The causal DAG for the RDD and our variant of it.

We now want to consider a scenario inspired by the RDD, which allows to bound the back-door dependence in the sense of Section 3.6 and thus makes Theorem 7 applicable to estimate the causal effect $\partial_2\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]$, i.e. the differential effect of treatment on the treated. The scenario differs from the RDD in that neither a discontinuity in the structural equation for X, nor linearity of the remaining structural equations is required.

Assume the causal DAG in Figure 3. Furthermore assume that $X = f_X(Z)$ for a function f_X that is differentiable. (This is the point where our scenario differs from RDD.) Suppose f_X is invertible, $g := f_X^{-1}$. It can easily be seen that this implies

$$\begin{aligned}&\partial_1\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]\\&= \partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]g'(x).\end{aligned}$$

Note that $\partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]$ means the derivative of $\mathbb{E}[Y|Z{=}z, \mathrm{do}\,X{=}x]$ w.r.t. z at position $(g(x), x)$. Applying Theorem 7 yields

$$\begin{aligned}&\mathrm{d}_x\mathbb{E}[Y|X{=}x] - \partial_2\mathbb{E}[Y|X{=}x, \mathrm{do}\,X{=}x]\\&= \partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]g'(x).\end{aligned}$$

Hence if for any position x_0 of X we assume a bound on the strength of the "back-door" dependence of Y on Z, $\partial_1\mathbb{E}[Y|Z{=}g(x_0), \mathrm{do}\,X{=}x_0]$, and if $|g'(x_0)|$ is comparably small (which is the case when $|f'_X(g(x_0))|$ is big), then we can bound the difference between observed dependence and causal effect at position x_0.

For instance, if we consider the observed dependence $\mathrm{d}_x\mathbb{E}[Y|X{=}x]$ as a realistic scale based on which one can constrain $\partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]$, formally

$$|\partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]| \leq c|\mathrm{d}_x\mathbb{E}[Y|X{=}x]|,$$

for some c, then one can bound the modulus of the causal effect from below:

$$\begin{aligned}&|\partial_2\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]|\\&\geq (1 - c|g'(x)|)|\mathrm{d}_x\mathbb{E}[Y|X{=}x]|.\end{aligned}$$

Obviously one weakness of the above argument is that the estimation of the causal effect heavily depends on the bound that we assume w.r.t. the "back-door" dependence of Y on Z, $\partial_1\mathbb{E}[Y|Z{=}g(x), \mathrm{do}\,X{=}x]$. However, this can be seen as a quantitative analogon to the qualitative assumption of the RDD that $\mathbb{E}[U|Z = z]$ is continuous in z.

Keep in mind that our results on Fisher information (Section 3.4) can be used in the case where X is not a deterministic function of Z that changes rapidly but instead the conditional probability $p(X|z)$ changes fast at some $z = z_0$.

5 CONCLUSIONS

In this paper, we analyzed a simple intuition linking observation and causation: if the observed dependence is strong and the effect of confounding is known to be weak, then we can infer a causal effect. We did this by employing a number of different notions for measuring dependence and causation, leading to different theoretical bounds. We do not argue that at present, there is a single formalization that best captures all aspects of this intuition, rather, we try to shed light on properties of the various notions by applying them to the same fundamental problem. While bounding confounding appears easier based on information theoretic quantities, expressing the influence from the treatment to the outcome variable by e.g. the effect of treatment on the treated (ETT) seems more relevant for practical purposes. We discussed several prototypical scenarios where bounds on confounding can be derived.

References

J. Angrist, G. Imbens, and D. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the International Joint Conference in Artificial Intelligence*, pages 357–363, Edinburgh, Scotland, 2005.

N. Ay and D. Krakauer. Geometric robustness and biological networks. *Theory in Biosciences*, 125:93–121, 2007.

N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–41, 2008.

T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.

B. Efron and D. Feldman. Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

G. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635, 2008.

D. Janzing, E. Sgouritsa, O. Stegle, P. Peters, and B. Schölkopf. Detecting low-complexity unobserved causes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Annals of Statistics*, 41(5):2324–2358, 2013.

D. Lee and T. Lemieux. Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48:281–355, 2010.

J. Pearl. *Causality*. Cambridge University Press, 2000.

J. Pearl. Direct and indirect effects. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 411–420, San Francisco, CA, 2001. Morgan Kaufmann.

J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

R. C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945. ISSN 0008-0659.

T. Richardson and J. Robins. Single world intervention graphs (swigs). Technical report, University of Washington, 2013.

J. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

J. Robins, T. VanderWeele, and T. Richardson. Discussion of ”causal effects in the presence of non compliance a latent variable interpretation” by forcina, a. *Metron*, LXIV(3):288–298, 2007.

D. Thistlewaite and D. Campbell. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51:309–317, 1960.