On Identification and Inference for Direct Effects

James Robins
Harvard University

Thomas Richardson
University of Washington

Peter Spirtes
Carnegie-Mellon University

November 26, 2009

Consider the query: *Does a binary treatment X have a causal effect on a response Y through a causal pathway that does not involve the intermediate variable M*? This query is often rephrased as: *Does X have a direct causal effect on Y not through M?* Direct effects have been formally defined in three different ways: the controlled direct effects (CDE), the natural direct effects (i.e. pure and total direct effects - PDE and TDE), and the principal stratum direct effects (PSDE). In this issue of the journal, Hafeman and VanderWeele (H&V) [7] provide novel minimal or near minimal conditions for identification of the CDE, PDE and TDE but do not consider the PSDE. In this commentary, we review inference for direct effects and the results of H&V. We also review the close relationship between the direct effects literature and the literature on instrumental variables and Mendelian randomization.

1 Formal Definitions

To proceed, we review the formal definitions of the three types of direct effects. We first consider a study with baseline covariates C, a dichotomous treatment X measured at start of follow-up, a binary intermediate M measured, say, at one month, and a binary outcome Y measured, say, at three months. Until Section 3.4 we focus on issues particular to identification of direct effects by following H&V and assuming (i) X is randomized conditional on C and (ii) at each level of C, the study population is sufficiently large that we can ignore sampling variability. Henceforth, unless stated otherwise, we restrict the population to the N subjects at a single joint level of C, and drop C from the notation. We use capital letters for random variables (i.e. variables whose value differs among subjects) and lower case letters for their possible values. Hence y, m, x can take the values 0 or 1.

The distribution of (X, M, Y) is fully characterized by the randomization probability $P\left(X=1\right)$ and the joint probabilities $P\left(Y=y, M=m \mid X=x\right)$ of Y and M within levels of X. We let M_x be a subject's counterfactual intermediate response when X is set to x and $Y_{x,m}$ be the counterfactual outcome when M is set to m and X to x. The counterfactual outcome Y_x when only X is set to x is, by definition, Y_{x,M_x}, i.e. $Y_{x,m}$ with m equal to the counterfactual intermediate M_x.

Our assumption that X was randomly assigned implies that X is jointly independent of the above counterfactuals, i.e.

$$\{Y_{x,m}, M_x; \text{for all } x, m\} \perp\!\!\!\perp X. \tag{1}$$

The observed variables M and Y are by definition M_x and Y_x evaluated at $x = X$. We say a causal effect is identified if it can be computed from the distribution of the observed data. Randomization of X implies that the population average causal effect of X on M

$$E\left[M_1 - M_0\right] = P\left[M_1 = 1\right] - P\left[M_0 = 1\right])$$

and X on Y

$$E\left[Y_1 - Y_0\right] = P\left[Y_1 = 1\right] - P\left[Y_0 = 1\right]$$

are identified by the risk differences $E\left[M \mid X=1\right] - E\left[M \mid X=0\right]$ and $E[Y \mid X=1] - E[Y \mid X=0]$ respectively.

The three types of direct effects are defined as follows. The two (average) CDEs are CDE$(m) \equiv E\left[Y_{1,m} - Y_{0,m}\right]$, $m = 0$ or $m = 1$. CDE(m) is the average effect of X on Y in the study population were, contrary to fact, all subjects to have Mset to m. The (average) pure and total direct effects are PDE $\equiv E\left[Y_{1,M_0} - Y_{0,M_0}\right] = E\left[Y_{1,M_0}\right] - E\left[Y_0\right]$ and TDE $\equiv E\left[Y_{1,M_1} - Y_{0,M_1}\right] = E\left[Y_1\right] - E\left[Y_{0,M_1}\right]$. The PDE measures the effect of X on Y when Z is set to its value $Z\left(x=0\right)$ under non-exposure to X. The TDE measures the effect of X on Y when Z is set to its value $Z\left(x=1\right)$ under exposure to X. The two principal stratum direct effects

$$\text{PSDE}\left(m\right) \equiv E\left[Y_{1,m} - Y_{0,m} \mid M_1 = M_0 = m\right] = E\left[Y_1 - Y_0 \ \mid M_1 = M_0 = m\right]$$

are the effect of X on Y among the subset of the population with the intermediate value m under both treatment with $X = 1$ and $X = 0$. None of these direct effect parameters are identifiable under the sole assumption that X was randomly assigned.

Prior to 1986, the three different kinds of direct effects were not distinguished from one another even though direct effects were widely discussed in the sociology literature. Causal inference in sociology was largely based on linear structural equation models (LSEMs) with additive effects and normally distributed, additive errors. In an LSEM, the individual direct effect $Y_{1m} - Y_{0m}$ is a constant β that does not depend on either the level m of the intermediate or on the subject; as a consequence, under a LSEM, all types of direct effects are identical and equal to β. In contrast to sociology, the data in biostatistical and epidemiological settings are frequently categorical or censored failure times, so nonlinear effects with interactions could not be avoided. The effects CDE(m) were explicitly defined in Robins[32]; see also Holland[8]. In Sec 12.2 of the same paper, Robins introduced PSDEs to define the causal effect of the treatment X on a specific cause of death, in the presence of censoring by competing causes of death. Rubin[35,36] and Frangakis and Rubin[3] later used the same PSDE contrast to solve the same problem of 'censoring by death' and provided the 'principal stratum' appellation. Robins and Greenland[26] introduced the PDE and TDE and the associated pure and total indirect effects. Pearl[19] referred to the PDE and TDE as natural direct effects.

In Section 2, following H&V and much of the literature we assume that the intermediate is binary. However in many settings the intermediate of scientific interest is continuous. In Sections 3.1 and 3.2 we argue that in this case one should refrain from replacing the continuous intermediate with a dichotomized version in the analyses. In fact, we show that the direct causal effect of X on Y not through the dichotomized version is not well-defined. We also argue that when the intermediate is continuous the principal stratum approach to direct effects is of little use. In section 3.3 we discuss the implications of our results concerning continuous intermediates for instrumental variable analyses and for Mendelian randomization. Finally in section 3.4 we consider a novel approach to the analysis of Mendelian randomization studies based on inverse probability of treatment weighting.

Throughout the paper we will use the notation $A \perp\!\!\!\perp B \mid C$ to indicate that A is independent of B within levels of C.

2 Identification and Inference for Direct Effects

Consider a) H&V's M-monotonicity assumption that X is never protective for M, so subjects for whom M would be zero when treated but one when untreated (i.e. $M_1 = 0, M_0 = 1$) do not exist and b) the assumption that $M_1 \perp\!\!\!\perp Y_{00} \mid M_0 = 0$, i.e.

$$\mathrm{OR}_0 = 1,$$

where we define OR_0 to be the conditional odds ratio between the counterfactuals M_1 and Y_{00} among subjects with $M_0 = 0$. The counterfactual association parameter OR_0 was first introduced into the literature by Jemiai *et al.*[11] in their recent paper on the estimation of PSDE(0). In Lemma 1 of the appendix, we show that when M-monotonicity holds and X is randomized, the assumption $\mathrm{OR}_0 = 1$ is equivalent to H&V's assumption 1 in their Table 4a that X and Y_{00}

are conditionally independent given $M = 0$ i.e. $X \perp\!\!\!\perp Y_{00} \mid M = 0$.

The assumption $\mathrm{OR}_0 = 1$ plus M-monotonicity and randomization of X suffices to identify PSDE(0) by the m-specific risk difference

$$\mathrm{RD}\,(m) = E[Y|X = 1, M = m] - E[Y|X = 0, M = m] \tag{2}$$

for $m = 0$. In fact, this result is a special case of a more general result of Jemiai *et al.*[11], who show that under M-monotonicity and randomization of X, PSDE(0) is identified if OR_0 is known. Furthermore they prove that the identifying formula for, and the value of, PSDE(0) varies with the true OR_0 and equals the RD(0) only when $\mathrm{OR}_0 = 1$. In Lemma 2 in the appendix we provide their identifying formula for PSDE(0). Finally they show that the true OR_0 itself is not identified and indeed the distribution of the observed data places no bounds on its value. A non-identified parameter, such as OR_0, that encodes the magnitude of *selection bias* or *non-comparability* and (were it known) suffices to identify a causal effect is often referred to as a selection bias parameter. Note that among subjects with $M_0 = 0$, the group with $M_1 = 1$ is *comparable* to the group with $M_1 = 0$ with respect to the distribution of the counterfactual Y_{00} if and only if $\mathrm{OR}_0 = 1$.

Jemiai *et al.*[11] saw no reason to privilege any particular value, such as 1, of the unidentified selection bias parameter OR_0. They, therefore, gave up on identifying PSDE(0) and, instead, conducted a sensitivity analysis in which they plotted PSDE(0) as a function of the true, but unknown and non-identifiable selection bias parameter OR_0. Except for those rare cases in which one strongly believes that the causal structure or mechanisms generating the data imply $\mathrm{OR}_0 = 1$, we agree that a sensitivity analysis approach is best (although, owing to the non-collapsibility of odds ratio measures, it is generally preferable to choose a selection bias parameter that quantifies the degree of selection bias or confounding on a additive or multiplicative, rather than on an odds ratio, scale[25]). Sensitivity analysis methodologies have been proposed for the CDE(m) by Robins, Rotnitzky & Scharfstein[27], and for the CDE(m), PDE, and TDE by VanderWeele[38]; see also Glynn and Quinn[6].

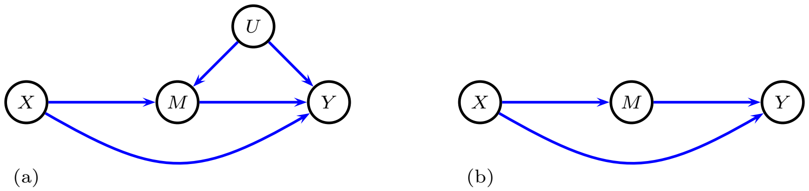

Figure 1: Graphs depicting the causal relationship between exposure (X), the mediator (M), and the response (Y), (a) when M is not randomized; (b) when M is randomized conditional on X.

To emphasize the importance of refraining from assuming that $\mathrm{OR}_0 = 1$ and thus that RD(0)=PSDE(0) without strong justification, we work out in Lemma A3 just how biased RD(0) can be when $\mathrm{OR}_0 \neq 1$. Specifically, we show that when X is randomized and M-monotonicity holds, the true PSDE(0) can be zero even when $RD\,(0)$ is as small as

$$-\min\left\{P(Y\!=\!1 \mid M\!=\!0, X\!=\!0), \left(\frac{P(M\!=\!0 \mid X\!=\!0)}{P(M\!=\!0 \mid X\!=\!1)} - 1\right) P(Y\!=\!0 \mid M\!=\!0, X\!=\!0)\right\}$$

or as large as

$$\min\left\{P(Y\!=\!0 \mid M\!=\!0, X\!=\!0), \left(\frac{P(M\!=\!0 \mid X\!=\!0)}{P(M\!=\!0 \mid X\!=\!1)} - 1\right) P(Y\!=\!1 \mid M\!=\!0, X\!=\!0)\right\}.$$

For example, suppose in our data, $P(M\!=\!0 \mid X\!=\!0) = 0.9$ and $P(M\!=\!0 \mid X\!=\!1) = 0.08$, $P(Y = 1 \mid M = 0, X = 1) = 1$, and $P(Y = 1 \mid M = 0, X = 0) = 0.1$ so $RD\,(0) = 0.9$. Then the true value of PSDE(0) could be zero, as the last display equals $\min\left(0.9, \frac{0.8}{0.08}0.1\right) = 0.9$, and thus does not exceed RD(0).

2.1 Joint 'randomization' of X and M

Causal DAGS, such as Figure 1(a), are commonly used to represent beliefs about causal structure or mechanism. In Figure 1(a), there are no common causes of X and M or of X and Y, as X was randomized. Since M was not randomly assigned by the investigator, M and Y may have a common cause U.

Suppose, however, one believed that 'nature' effectively randomly assigned M, with randomization probabilities that (possibly) depended on the randomized treatment X in the sense that

$$\{Y_{x,m}; \text{all } x, m\} \perp\!\!\!\perp M \mid X. \tag{3}$$

Then M and Y would have no unmeasured common cause and the causal graph would be that in Fig. 1(b). Further X and M (since jointly randomized) would be jointly independent of all the $Y_{x,m}$. This joint randomization of M and X, formalized by the conjunction of Eqs. (1) and (3), implies that for all x, x^{*}, m in $\{0,1\}$,

$$\begin{aligned} (Y_{x,m}, M_{x^{*}}) &\perp\!\!\!\perp X \\ Y_{x,m} &\perp\!\!\!\perp M \mid X \end{aligned}$$

which are H&V's Eqs. A1 and A6. In their appendix H&V refer to this latter conjunction as 'sequential ignorability'. We adopt their nomenclature, although their definition of 'sequential ignorability' does not exactly agree with previously published definitions.

In fact, under joint randomization of M and X and M-monotonicity, all direct effects are identified: PSDE(m) and CDE(m) both by RD(m), TDE and PDE by sRD(1) and sRD(0) respectively, where

$$\text{sRD}\left(x\right) \equiv \sum_{m} \text{RD}\left(m\right) P\left(M = m \mid X = x\right).$$

We prove this result for the PSDE(1) in Lemma 5 in the appendix. For PSDE(0) the result follows from the fact that joint randomization of M and X implies $X \perp\!\!\!\perp Y_{00} \mid M = 0$ which by Lemma 1 implies $\text{OR}_0 = 1$.

For the other direct effect parameters M-monotonicity plus the assumption of sequential ignorability, which is weaker than Eqs. (1) and (3), suffices for identification. This follows from the fact that under sequential ignorability and M-monotonicity, the assumptions mentioned in H&V's Tables 4a-4d and Eqs A.1-A.7, including those ascribed to Pearl, Kaufman[12], Robins and Greenland[26], Imai *et al.*[10], and Petersen, Sinisi and van der Laan[21], all hold.

2.2 Identification with joint 'randomization' but without M-monotonicity

Robins (ref. 32; ref. 33, page 76) noted the above identifying formulae are true for the CDE(m), PDE, and TDE under joint randomization of M and X, even without assuming M-monotonicity. This result also follows from Imai *et al.*'s[10] and H&V's results on 'sequential ignorability'.

In fact it is easy to show that sequential ignorability, even in the absence of M-monotonicity, implies the assumptions mentioned in H&V's Tables 4a-4d and Eqs A.1-A.7 provided that, in Table 4c, the actual assumption

$$E\left[Y_{1m} - Y_{0m} \mid M_0 = 1\right] = E\left[Y_{1m} - Y_{0m} \mid M_0 = 0\right] \tag{4}$$

made by Petersen, Sinisi and van der Laan is used instead of H&V's assumption that $P\left[Y_{1m} - Y_{0m} = 1 | M_0 = 1\right] = P\left[Y_{1m} - Y_{0m} = 1 \mid M_0 = 0\right]$, which is equivalent to (4) only under M-monotonicity.

In contrast, joint randomization of M and X, i.e. the conjunction of Eqs. (1) and (3), is not sufficient to identify PSDE(m) in the absence of M-monotonicity. However, the PSDE(m) is identified by the RD(m) if Eq. (3) is strengthened to

$$\{Y_{x,m}; \text{all } x, m\} \perp\!\!\!\perp (M_0, M_1) \mid X \tag{5}$$

which essentially says that, within levels of X, nature randomly assigned (M_0, M_1) by rolling a 4 sided die; see Lemma 4.

Eq. (5) implies Eq. (3). Thus the counterfactual causal model defined by Eqs. (1) and (5) encodes stronger (i.e. more restrictive) causal assumptions than the model defined by Eqs. (1) and (3). This is the reason that, in the absence of M-monotonicity, the PSDE(m) is identified under the former but not the latter; in contrast, as noted above, the CDE(m), PDE, and TDE are identified under both models.

The reader may well find the results of the previous paragraph surprising or puzzling for the following reason. In the epidemiologic literature on causal DAGs, one often sees the statement that a causal DAG represents a particular causal structure. For example, the DAG of Figure 1(b) is said to represent a structure in which there is no confounding between M and Y given X. However, to uniquely and formally define the causal structure associated with a DAG, we must associate with the DAG a given set of counterfactual conditional independence statements. Yet both Eq. (5) and Eq. (3) seem a reasonable way to encode the absence of confounding between M and Y given X. Thus, we must accept that we can associate a single graph with somewhat different counterfactual conditional independences. We might hope that these differences do not lead to different conclusions. Unfortunately, this is not the case. For example, as described above, in the absence of M-monotonicity, PSDE(m) is identified under Eq. (5), but not under Eq. (3).

Elsewhere[33,22] we argue at length that such different conclusions concerning identification can only occur with causal effects such as the PSDE(m), PDE(m), and TDE(m) that, unlike the CDE(m), are not 'manipulative' effects. A manipulative effect is, by definition, the causal effect of an intervention on (i.e. a manipulation or setting of) some of the graph variables to known values (so the intervention could, in principle, actually be conducted) on an identified subset of the population. For example, the PSDE(m) represents the effect on Y of setting X to 1 versus 0 on a non-identified subset of the population - the unknown subjects whose intermediate takes the value m regardless of their X treatment, hence the PSDE(m) is a not a manipulative parameter. Likewise, PDE(m) is the effect on the entire population of setting X to 1 versus 0, while always setting M to the value it would have under $X = 0$. However the intervention that sets X to 1 and M to the value it would have under $X = 0$ cannot be conducted, even in principle, because, once X is set to 1, the value of M under $X = 0$ will be unknown. Similar comments apply to the TDE.

In the remainder of this section, we will briefly outline some of the different approaches that have been proposed for associating counterfactual independence relations with a causal DAG. For a more detailed discussion, see References 33 and 22. *Some of the issues raised in our outline are subtle and technical and may be skimmed over on a first reading; §2.3 onwards may be read independently* *of this material.*

Counterfactual frameworks

A *counterfactual framework* is a rule for associating a set of counterfactual conditional independence relations with a given causal DAG; these are in addition to the conditional independence relations that follow from d-separation applied to the DAG. A given DAG together with the counterfactual conditional independence relations implied by a framework define a framework-specific causal model; this model is said to be associated with the DAG (under a given framework). Though several such frameworks have been described, we focus on two:

The non-parametric structural equation (NPSE) framework of Pearl (See Ref. 17, p.101) is the strongest or most restrictive counterfactual framework, in the sense that, relative to other frameworks, it associates the largest set of counterfactual conditional independence relations with a given DAG. Consequently, more quantities are identified in this framework than in any other. The NPSE model (NPSEM) associated with the causal DAG in Figure 1(b) is equivalent to the counterfactual model defined by Eqs. (1) and (5) and thus the NPSEM identifies the PSDE(m), PDE(m), and TDE(m).

In contrast, Robins and Richardson[22] describe a minimal counterfactual framework, which is the least restrictive framework under which all intervention distributions are identified; see Appendix B for the independence relations corresponding to the graph in Figure 1(b). Robins and Richardson show that the PDE, TDE and TDE are not identified under minimal counterfactual model (MCM) associated with the causal DAG in Figure 1(b), although the CDE(m) remains identified. It follows that the MCM is less restrictive than even the 'sequential ignorability' model of H&V, as the latter sufficed to identify the PDE in our example.

Plausibility of different ‘randomization’ assumptions

The existence of these alternative frameworks naturally raises the question as to when it is reasonable to either believe or disbelieve that ‘nature’ effectively randomly assigned M within levels of X (and baseline covariates C), either in the strong sense that Eq. (5) and thus the NPSEM associated with Figure 1(b) holds or in the weak sense that the MCM associated with Figure 1(b) holds.

If M is an action performed by a human agent it is often possible to have a reasoned opinion on the matter. As an example, suppose the randomized treatments $X = 1$ and $X = 0$ denote new and standard beta-agonist drugs in a trial with the outcome Y denoting hospitalization in the third month for asthma, and M denoting nonrandomized treatment with oral corticosteroids at 1 month. Since doctors prescribe oral corticosteroids for worsening asthma, the worsening of asthma over the first month would be an unrecorded common cause U of M and Y, ruling out the DAG in Fig. 1(b) and thus both any associated MC and/or NPSEM framework.

If M is a biological intermediate, a reasoned opinion would require a level of biological knowledge that often does not exist. For example in Jemiai *et al.*, X was a trial HIV vaccine, M denoted infection with HIV, and Y was a measure of HIV progression (e.g. CD4 count). The question of the existence of U in Fig. 1(a) is then the question of whether there exist unmeasured (genetic or environmental) factors that determine, in part, both infection with HIV and the rate of disease progression once infected, within joint levels of measured baseline clinical, laboratory, and life-style factors C. A remarkable development of the last several years is that evidence concerning such genetic factors is beginning to be obtained from GWAS studies.

In general, it would be rare to believe strongly that M was ‘randomized’ in the strong or even the weak sense. On those rare occasions where weak sense belief is justified, strong sense belief will generally also be justified, as causal processes that satisfy weak but not strong sense tend to require special, substantively unlikely, tuning. However, in the unlikely event that ‘nature’ randomized M only in the weak sense, yet inference was conducted under the assumption that 'nature' had randomized in the strong sense holds, our estimates of PSDE(m), PDE and TDE will be biased. This would be particularly unfortunate because, in the absence of additional strong background knowledge, owing to the fact that the PSDE(m), PDE and TDE are non-manipulative effects, this bias could not be detected, even were unlimited resources available to conduct designed experiments wherein the investigators randomly assigned one or both of X and M (under any randomization scheme).

2.3 An MSC Model in which PSDE(0) is identified but sequential ignorability does not hold

As discussed above, when M is 'randomized by nature' in the strong sense, all previously proposed identification conditions for direct effects based on data (X, M, Y) are true. Viewed in this light, we believe the primary contribution of H&V is to propose a novel causal structure in which there exists an unrecorded common cause U of M and Y under which X and Y_{00} are independent given $M = 0$ hold and thus, in which if M-monotonicity holds, PSDE(0) is identified. In this section, we describe their causal structure.

In Figure 1(a), X and U are independent causes of M. This implies that X and U must be conditionally dependent in at least one of the two strata of M. However, H&V show that X and U (called G by H&V) remain conditionally independent within the stratum $M = 0$ under a minimal sufficient cause model (MSC) in which X and U act through separate independent mechanisms to cause $M = 1$. It then follows that X and Y_{00} are also independent given $M = 0$, as we demonstrate graphically below. Hence, given M-monotonicity, PSDE(0) is identified by RD(0).

Figure 2 is a causal DAG that has added to the DAG in Figure 1(a) an MSC model for M (that includes H&V's model as a special case) where M^a, M^b, and M^c are possibly unobserved, binary variables encoding sufficient causes of M and $Y_{\mathbf{x},\mathbf{m}}$ is the vector of the 4 counterfactuals $Y_{x,m}$. The separate mechanisms

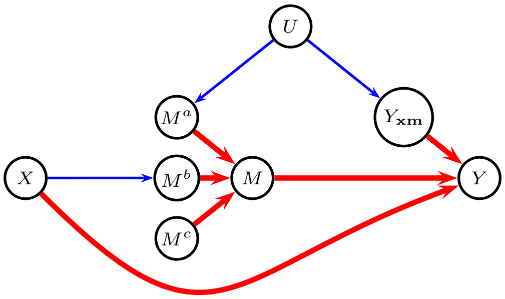

Figure 2: A DAG showing an underlying MSC model, under which $X \perp\!\!\!\perp Y_{00} \mid M = 0$. Red edges indicate deterministic relationships.

by which X and U, respectively, cause M are encoded in (i) the arrows from U to M^a, X to M^b and (ii) in the bold arrows from M^a, M^b and M^c to M. These bold arrows indicate that M is a deterministic function of its parents, in particular we specify that M occurs (i.e. takes the value 1) if and only if one or more of the sufficient causes M^a, M^b, or M^c occurs (equals 1). The independence of the mechanisms is encoded by M^a, M^b and M^c having no common causes. It follows that, whenever we condition on $M = 0$, we also condition on $M^a = M^b = M^c = 0$. Hence since X and $Y_{\mathbf{x},\mathbf{m}}$ are d-separated given $\{M^a, M^b, M^c\}$, we conclude that the vector $Y_{\mathbf{x},\mathbf{m}}$ is independent of X given $M = 0$. (See VanderWeele and Robins[39] for additional discussion.)

Although H&V's MSC model identifies PSDE(0) under M-monotonicity, we now show that it fails to identify any of the other direct effects, except when the direct effect sharp null hypothesis

$$Y_{1m} - Y_{0m} = 0, \text{ for all } m \text{ and all subjects} \tag{6}$$

holds so all direct effects are equal to zero. Graphically, the null hypothesis (6) is captured by removing the direct arrow from X to Y in Figures 1 and 2. To identify PSDE(1), we would need X and U to be conditionally independent given $M = 1$, but, as argued above, this is not possible when X and U are conditionally independent given $M = 0$. Furthermore, under H&V's MSC model, one can check using d-separation applied to Figure 2, that none of the (joint) assumptions of H&V, Pearl, Kaufman, Robins and Greenland, Van der Laan and Petersen, and Imai needed to identify CDE(m), PDE or TDE are satisfied, whenever the arrow from X to Y is present on the DAG in Fig. 1(a) (i.e. whenever the direct effect sharp null hypothesis (6) is false). This result does not contradict H&V, since, in their example, H&V assumed the direct effect sharp null hypothesis holds. In that case, all direct effects are zero. Consequently, since RD(0) is consistent for PSDE(0) $= 0$, it is also consistent for all the other direct effects including CDE($m\!=\!0$), as noted by H&V.

Finally, by an analogous argument, a minimal sufficient cause model (MSC) in which X and U act through separate independent mechanisms to cause $M = 0$ (rather than $M = 1$) identifies PSDE(1), but fails to identify PSDE(0) or any of the other direct effects, except when the direct effect sharp null hypothesis (6) holds.

In summary, although these results on identification in an MSC model are of theoretical interest, their practical utility is constrained both by the fact that, unless the direct effect null hypothesis (6) holds, the only direct effect parameter identified is the PSDE(0) and, more importantly, by the fact that it would be rare for one to hold a strong belief that the data have been generated by a specific causal model.

Thus, as discussed above, we believe the best inferential approach for all 3 types of direct effect parameters would usually be to display a sensitivity analysis in which the direct effect parameter is plotted as a function of a non-identifiable selection bias parameter.

3 Continuous Intermediates, Principal Stratum Direct Effects, and Instrumental Variables.

Much of the literature on direct effects assumes a binary intermediate. However, the binary intermediate, say M^*, is often constructed by dichotomizing an ordinal or continuous intermediate, say M, at a cut-point. For example, in Pearl[18], the intermediate M was the fraction of the prescribed pills that were actually consumed in the Lipid Research Clinics Primary Prevention Trial and the cut-point was chosen to be halfway between the minimum and maximum fraction. Similarly, Kaufman *et al.*[13] and Cai *et al.*[1] examined the direct effect of cholestyramine on coronary heart disease, not mediated via serum cholesterol, using a dichotomized version of serum cholesterol as the intermediate.

In the next subsection, we shall assume that, as is true in most analyses of direct effects, the question of scientific interest is whether the treatment X has a direct causal effect on a response Y through a causal pathway that does not involve the continuous intermediate M, i.e. whether the direct effect sharp null hypothesis (6) holds. We will argue that (i) one should refrain from replacing M by its dichotomized version $M^* \equiv I(M \geq \zeta)$, for some cut-off ζ (here $I(\cdot)$ is the indicator function), and (ii) the principal stratum approach to direct effects is of little utility, regardless of whether the principal strata are defined in terms of M or M^*. In subsection 3.2, we consider the question of whether the treatment X has a direct causal effect on a response Y through a causal pathway that does not involve the dichotomized intermediate M^*. Finally in subsections 3.3 and 3.4, we discuss the implications of our results for instrumental variable (IV) methods and Mendelian randomization.

3.1 Direct effects not through M

Suppose that M is a cause of Y, and 'nature' effectively randomly assigned M in the strong sense that Eqs. (1) and Eq. (5) hold, so that the NPSEM associated with the causal DAG in Fig. 1(b) is the data generating mechanism. Then, PDE, TDE and CDE(m) are identified by sRD(0), sRD(1), and RD(m) , just as in the M binary case. [If M is continuous an integral replaces the sum in the definition of sRD(x).] In particular under the null hypothesis (6), sRD(0), sRD(1), and RD(m) are all zero. In contrast, one cannot use the data (X, M^*, Y) to validly test the null hypothesis (6) of scientific interest, since, under the null (6), the

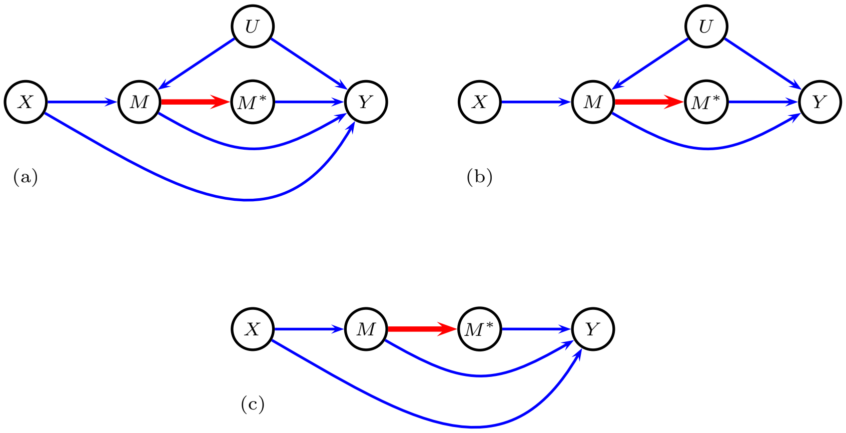

Figure 3: DAGs in which M^* is a dichotomized version of M; (a) with a direct effect of X on Y; (b) the no direct effect (relative to M) sharp null hypothesis; (c) no confounding. As before, red edges indicate deterministic relationships.

contrasts

$$\mathrm{RD}^*\left(m^*\right) = E[Y|X=1, M^*=m^*] - E[Y|X=0, M^*=m^*], \quad m^*=0,1$$

and

$$\mathrm{sRD}^*\left(x\right) = \sum_{m^*} \mathrm{RD}\left(m^*\right) P\left(M^*=m^*|X=x\right)$$

based on the *dichotomized* exposure M^* will be non-zero, except if, as will very rarely be the case, the effect of M on Y is entirely through M^*, i.e. for each subject, the counterfactual outcome $Y_{1,m} = Y_{0,m}$ under the null (6) is identical for all values of m on the same side of the cut-point ζ. To understand the above graphically, consider Figure 3. The DAG in Fig. 3(a) adds the variable M^* to Fig. 1(a) with the bolded arrow from M to M^* indicating the deterministic relationship given by the dichotomization. The DAG in Fig. 3(b) encodes the null hypothesis (6) by eliminating the arrow from $X \rightarrow Y$. The DAG in Fig. 3(c) additionally encodes the assumption that 'nature' effectively randomly assigned M by removing U. In any of these scenarios, if the effect of M on Y were entirely mediated through M^*, the edge $M \rightarrow Y$ would also be removed.

Consider now the PSDE(m). Recall PSDE(m) was identified in the binary M case, under this NPSEM. Suppose, however, that the intermediate M is now continuous and that X has an effect on every subject's M. In that case,

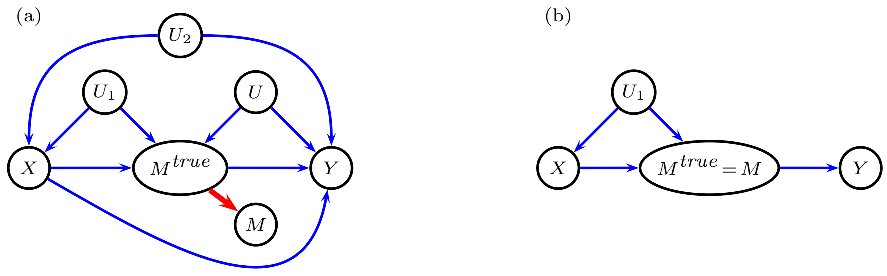

Figure 4: (a) A DAG in which M is a dichotomized version of M^{true}; (b) with a direct effect of X on Y. The DAG in (b) may be inferred via faithfulness when $X \perp\!\!\!\perp Y \mid M$ is observed to hold, having assumed the DAG in (a) initially.

there is no subject who has $M_1 = M_0 = m$ for any m. As a consequence, principal stratum direct effects PSDE(m) do not exist. When M is continuously distributed, the hypothesis that X has an effect on every subject's M cannot be empirically rejected. Thus the principal stratum approach to direct effects based on PSDE(m) is of little use for continuous M in the absence of further untestable assumptions.

In contrast to the PSDE(m), at least one, and often both, of the two principal stratum direct effects $\text{PSDE}^* (m^*) = E\left[Y_1 - Y_0 \ \mid M_1^* = M_0^* = m^*\right], m^* = 0, 1$ based on M^* will be well defined, where $M_x^* = 1$ if M_x exceeds the cut-off and $M_x^* = 0$ otherwise. However, even under the null hypothesis (6), $\text{PSDE}^* (m^*) = E\left[Y_1 - Y_0 \ |M_1^* = M_0^* = m^*\right] = E\left[Y_{1,M_1} - Y_{0,M_0} |M_1^* = M_0^* = m^*\right]$ will generally be non-zero and thus will be of little scientific interest[23], even were it identified. To see why, note that, under (6), $Y_{1,M_1} - Y_{0,M_0} = Y_{1,M_1} - Y_{1,M_0}$, but since even among the subset of subjects with $M_1^* = M_0^* = m^*$, the distribution of M_1 will differ from that of M_0, the means of Y_{1,M_1} and Y_{1,M_0} will still differ. Only when the effect of M on Y is entirely through M^*, as can be graphically represented by removing the $M \rightarrow Y$ arrow from the graph in Figure 3(b), will $\text{PSDE}^*(m^*) = 0$.

3.2 Direct effects not through M^*

In this subsection we examine the question of whether the treatment X has a direct causal effect on the response Y through a causal pathway that does not involve the dichotomized intermediate M^* under the sole assumption that Fig. 1(a) is the causal DAG associated with a counterfactual model, as we want to allow for confounding of the effect of M on Y. The natural interpretation of this question is that we would like to know whether the direct* effect sharp null hypothesis, i.e.

$$Y_{1m^*} - Y_{0m^*} = 0, \text{ for } m^* = 0,1 \text{ and all subjects} \tag{7}$$

holds, where Y_{xm^*} is the counterfactual value of Y when X is set to x and M^* is set to m^*. Graphically, we want to know whether any of the directed paths from X to Y not through M^* in Figure 3(a) are actually present.

However, technically we have to first ask whether the counterfactuals Y_{xm^*} exist as well-defined functions of the underlying counterfactuals M_x and Y_{xm} associated with the causal DAG of Fig. 1(a). Let ζ denote the cut-off value for dichotomization. Then $Y_{x,m^*=1}$ is naturally defined as $Y_x = Y_{x,M_x}$ when $M_x^* = 1$ or equivalently $M_x \geq \zeta$, but is not uniquely defined when $M_x^* = 0$. Similarly $Y_{x,m^*=0}$ is naturally defined as $Y_x = Y_{x,M_x}$ when $M_x \leq \zeta$, but is not uniquely defined when $M_x > \zeta$. Thus $Y_{1,m^*=1} - Y_{0,m^*=1}$ is only well-defined for subjects with $M_1^* = M_0^* = 1$ and $Y_{1,m^*=0} - Y_{0,m^*=0}$ is only well-defined for subjects with $M_1^* = M_0^* = 0$. Thus, absent further assumptions, the only meaningfully defined tests of the null hypothesis (7) are tests of the hypotheses that $\text{PSDE}^*(m^*) = E\left[Y_1 - Y_0 \mid M_1^* = M_0^* = m^*\right] = 0$ for $m^* = 0,1$.

Richardson and Robins[22] derived empirical tests of these latter hypotheses based on the data (X, M^*, Y) for X and Y binary under the NPSEM associated with Figure 1(a). Their tests do not assume M^*-monotonicity and are identical to the tests of the hypotheses $\text{CDE}^*(m^*) = E\left[Y_{1,m^*} - Y_{0,m^*}\right]$ equals 0, for $m^* = 0,1$ derived by Cai *et al.*[1] and Kaufman *et al.*[13] under the additional assumption that all the Y_{xm^*} were well-defined. (The same tests are obtained owing to the fact that, for subjects other than those with $M_1^* = M_0^* = 1$ and

$M_1^* = M_0^* = 0$, mere existence of the Y_{xm^*} does not place empirical constraints on the means of $Y_{1,m^*=0} - Y_{0,m^*=0}$ or $Y_{1,m^*=1} - Y_{0,m^*=1}$.) Furthermore, the tests are also the same as Pearl's 'instrumental inequality' tests[16] of the hypothesis (7).

When the counterfactuals $Y_{x,m^*=1}$ and $Y_{x,m^*=0}$ are not well-defined for all subjects, the DAGs in Figure 3 cannot be associated with any counterfactual models, neither the CDE* (m^*) nor the contrasts $\mathrm{E}[Y_{x,m^*=1} - Y_{x,m^*=0}]$ encoding the effect of M^* when X is set to x are well defined contrasts, and the graphical arguments of the previous subsection are not quite technically correct. One approach, that is perhaps most natural, to making $Y_{x,m^*=1}$ and $Y_{x,m^*=0}$ well-defined functions of the counterfactuals $Y_{x,m}$ and M_x is to define $Y_{x,m^*=1}$ (respectively, $Y_{x,m^*=0}$) to be $Y_{x,m}$ evaluated at the cut-point $m = \zeta$ when $M_x < 0$ (respectively, $M_x > 0$).

3.3 Instrumental Variable Analysis using Dichotomized Continuous Intermediates

Suppose we are considering performing an 'instrumental variable' (IV) analysis of the effect of a dichotomous intermediate M^* on a response Y using X as the instrument based on the data (X, M^*, Y). One can then test the null hypothesis (7) using the aforementioned tests, i.e. Pearl's 'instrumental inequality' tests. If (7) is rejected, then the exclusion restriction, required for an IV analysis, does not hold, X is not an instrument, and the analysis should not proceed. Formally, an IV analysis is a method for estimation of the contrast $\mathrm{E}[Y_{x,m^*=1} - Y_{x,m^*=0}]$ under the assumption that (7) holds. In other words, such an IV analysis presupposes that the instrument X has no effect on Y, other than through M^*, and exploits this assumption to make inferences about the effect of M^* on Y. In the IV context, this assumption is called the exclusion restriction (relative to M^*).

However, owing to the fact that neither the PSDE* (m^*) nor the CDE* (m^*) are identified under the causal DAG in Figure 3(a), it follows that, even when hypothesis (7) is not rejected, the null hypothesis (7) and thus the exclusion restriction can still be false, even with an enormous sample size. As a consequence, if, based on subject matter considerations, it is thought that the null hypothesis (7) is likely false, the appropriate decision is to forego an IV analysis that assumes X is an instrument for M^*, regardless of whether a test of the null hypothesis (7) does or does not reject; therefore, no test of (7) need be performed.

We now argue that when M^* is a dichotomized version of a continuous or ordinal M, the null hypothesis (7) will likely be false whenever X and Y are dependent (i.e. X has a causal effect on Y), and thus an IV analysis that assumes X is an instrument for M^* should not be performed in this setting. When X and Y are dependent, the null hypothesis (7) will be false, unless X has no direct effect on Y not through M (i.e. the DAG in Figure 3(b) generated the data) and the effect of M on Y is only through M^* (i.e. the edge $M \rightarrow Y$ could be removed from DAG in Fig. 3(b)). However, we argued earlier, that, when M has an effect on Y, it would be very rare for that effect to be entirely through M^*.

3.4 Mendelian Randomization, Instrumental Variables, and Faithfulness

Our discussion of the difficulties involved in the interpretation of instrumental variable analyses carry over to the interpretation of studies involving Mendelian randomization[15]. Lawler et al.[15] and Didelez and Sheehan[2] provide comprehensive overviews of Mendelian randomization. We briefly touch on several points, also raised by these authors.

Consider a prospective Mendelian randomization study whose goal is to determine whether C-Reactive Protein (CRP) is a cause of myocardial infarction (MI) among 50 year old men. Data are obtained on blood levels of CRP (M) at age 50, a continuous intermediate, on incident MI (Y) over the following 5 years, and on genetic variants in the gene coding for the CRP. We restrict considerations to wild-type and certain particular variant alleles that cause a decrease in the expression level of the CRP without effecting the amino acid sequence. For example the variant alleles might represent mutations in a promoter region physically adjacent to an exon. For simplicity, we denote homozygous wild type by $X = 1$ and all other genotypes by $X = 0$. Suppose that the study is sufficiently large that sampling variability can be ignored. Further, suppose we observe that the CRP gene is positively associated with both CRP levels and MI, i.e. $E[M|X=1] - E[M|X=0] > 0$ and $E[Y|X=1] - E[Y|X=0] > 0$. Since the genetic variants X are physically located in the CRP gene, we assume, based on our knowledge of biology, that X had no direct effect on Y other than through its gene product M, i.e. the exclusion restriction (6) holds. Further assume the $X - Y$ relationship is unconfounded (e.g., population stratification is absent).

Then X is an instrument and X and Y being marginally dependent implies that the CRP protein has a causal effect on Y.

If we observed that X and Y were also associated given M, so $E[Y \mid M = m, X = 1] \neq E[Y \mid M = m, X = 0]$, we would additionally conclude either a) that there exists an unmeasured common cause U of M and Y (such as ongoing sub-clinical inflammation) as in Figure 1(a) and conditioning on M opens a collider path and/or b) M was a mis-measured version of the biologically relevant dose, say M^{true}, of CRP.

Let $\beta = \{E[Y|X=1] - E[Y|X=0]\} / \{E[M|X=1] - E[M|X=0]\}$ be the usual instrumental variable estimand. The causal effect of M^{true} quantified by the response function $r(m^{true}) \equiv E[Y_{x,m^{true}} - Y_{x,m^{true}=0}]$ would be equal to βm^{true}, under the non-identifiable assumptions that (i) measurement error is absent (i.e. $M = M^{true}$), (ii) the linear structural mean model

$$E\left[Y_{x,m^{true}} - Y_{x,m^{true}=0} \mid M^{true} = m^{true}, X = x\right] = \alpha m^{true}$$

with no effect modification by the instrument X is true with slope parameter α, and (iii) the average effect of the intermediate in the subset with intermediate value m^{true} equals that in the subset with intermediate value not equal to m^{true}, i.e.

$$\begin{aligned} E\left[Y_{x,m^{true}} - Y_{x,m^{true}=0} \mid M^{true} = m^{true}\right] & \\ = \quad E\left[Y_{x,m^{true}} - Y_{x,m^{true}=0} \mid M^{true} \neq m^{true}\right] & \end{aligned}$$

In fact under (i) -(iii), the slope parameter α equals the instrumental variable estimand β. If we substituted the assumption that LSEM given by $Y_{x,m^{true}} = \alpha m^{true} + Y_{0,m^{true}=0}$ holds for assumptions (ii) and (iii), we could still conclude that $r(m^{true}) = \beta m^{true}$ because the LSEM implies (ii) and (iii), although the converse is false[24].

However, if we replaced assumptions (ii) and (iii) by the weaker assumption that "$r(m^{true}) = \alpha m^{true}$ is a linear function of m^{true} with a slope α," we could not conclude $r(m^{true}) = \beta m^{true}$; indeed the dose response function would not be identified.

In summary, under the assumption that X is a valid instrument, although we can conclude CRP protein has a causal effect on MI owing to the marginal dependence of X and Y, we cannot identify the true dose response function $r(m^{true})$ without additional non-identifiable assumptions such as (i)-(iii), one or more of which is likely to be incorrect.

The inability to identify the true response function $r(m^{true})$ is an important limitation of a Mendelian randomization analysis even when the genetic variable X is a valid instrument. As an extreme example, suppose (ii) and (iii) held but (i) was false because the biologically relevant dose M^{true} of CRP determining MI in adult life was the cumulative *in utero* dose; that is, adult CRP has no effect on MI. Then, not only would βm^{true} differ generally from $r(m^{true})$, but, more importantly, an analyst who naively assumed (i) was true would incorrectly conclude that an intervention targeting adult CRP levels would prevent MI.

The interpretation of Mendelian randomization studies becomes even more problematic when the intermediate M is not the protein product of the gene X. As an example consider a prospective Mendelian randomization study whose goal is to determine whether body mass index (BMI) is a cause of MI. Data are obtained on the continuous intermediate BMI (M), on incident MI (Y), and on genetic variants in the FTO gene, a known genetic cause of increased BMI. For simplicity, we define $X = 1$ if both a subject's FTO alleles are known to cause increased BMI; otherwise $X = 0$. Suppose we observe that the effect of the FTO gene on both BMI and MI is positive, i.e. $E[M \mid X = 1] - E[M \mid X = 0] > 0$ and $E[Y \mid X = 1] - E[Y \mid X = 0] > 0$. Nonetheless, even when X is unconfounded, we cannot conclude that BMI has a causal effect on MI, because we lack a strong biological argument as to why the exclusion restriction, i.e, the null hypothesis (6), should hold, as BMI is not the protein product of the FTO gene. Furthermore, whether the exclusion restriction holds cannot be empirically tested when M is a continuous variable such as BMI without further non-identifiable assumptions. Likewise, the assumption that there is no confounding of the $X - Y$ association by unmeasured factors cannot be empirically rejected.

Faithfulness

Is there then any possibility of using the study data to help determine whether BMI may cause MI? We consider one rather controversial approach[29,4,30,5,28] that could be useful in the CRP example as well. Spirtes, Glymour and Scheines[37] and Pearl and Verma[20] have shown under a particular untestable assumption, referred to as 'faithfulness', it is sometimes possible to simultaneously discover from the data that the exclusion restriction (6) holds, that X is unconfounded with Y, that measurement error in M is absent, and that M is unconfounded within levels of X and baseline covariates! Faithfulness is the assumption that all conditional and unconditional independences 'found' in the data are due to causal structure and not to balancing of different causal effects.

Specifically, in a Mendelian randomization study in which X is marginally associated with M and Y, it turns out that faithfulness allows additional causal conclusions to be drawn from the data on (X, M, Y), but only when X and Y are found to be conditionally independent given M and some (possibly not strict) subset C_{sub} of the baseline covariates C. To understand the surprising

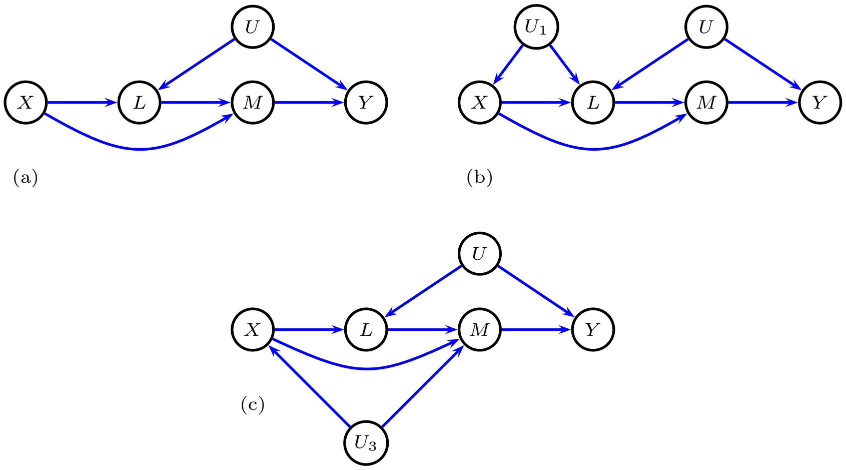

Figure 5: Three DAGs which imply $X \perp\!\!\!\perp Y \mid M$ in the pseudo-population created by re-weighting via $f^{pseudo}(x, l, m, y) \equiv f(x, l, m, y)\,(f(m)/f(m \mid x, l))$.

implications of assuming faithfulness, suppose the causal DAG in Figure 4(a) represented our *a priori* state of knowledge (within a level of C_{sub}). Figure 4(a) differs from Figure 1(a) in that we allow for the realistic possibility that X may be confounded (say, due to population substructure) and that M may be a randomly mis-measured version of the biologically relevant measure M^{true} of BMI. If we 'observe' that X and Y are independent given M, then, as explained next, faithfulness allows us to conclude that the true causal graph is that in Figure 4(b): That is, within levels of C_{sub} (i) M is the biologically relevant measure of BMI, (ii) M and Y have no unmeasured common cause U, (iii) X and Y have no common parent U_2, (iii) the direct arrow from X to Y is not present so the exclusion restriction (6) holds, and (iv) the true dose response function $r(m) = E\left[Y_{x,m} - Y_{x,m=0}\right]$ is identified and given by $E\left[Y|M = m\right] - E\left[Y|M = 0\right]$, as X is not a confounder for M in Figure 4(b).

These inferences may be obtained as follows: First, since $E\left[M|X = 1\right] - E\left[M|X = 0\right] > 0$, either U_1 must be a common cause of X and M^{true} or the $X \rightarrow M^{true}$ edge must be present, or both (so that an unblocked path exists from X to M). It follows, by faithfulness, we can remove U, U_2, and the $X \rightarrow Y$ edge from Figure 4(a), as X and Y would be dependent given M if either (i) U or U_2 were present as an unmeasured common cause or (ii) if the $X \rightarrow Y$ edge were present. It then follows from $E\left[Y|X = 1\right] - E\left[Y|X = 0\right] > 0$ that the $M^{true} \rightarrow Y$ edge must be present (so that there is a path d-connecting X and Y unconditionally). Next, by faithfulness again, M^{true} must equal M, as otherwise X and Y would be dependent given M. Thus we are left with Figure 4(b) and conclude M is M^{true}.

It follows that, if one is willing to invoke faithfulness, effort should be expended to try to make X and Y independent given M within levels of the measured baseline covariates. For example, one could try to make M closely approximate M^{true} by measuring the intermediate accurately and repeatedly over time, and then defining M to be the function of these repeated measurements that best predicts the outcome Y. Second, one could attempt to measure and add to the baseline covariates C, various hypothesized common causes, say L, of M and Y and X and Y, where L, unlike C, can be temporally subsequent to X (i.e. to conception).

Extending faithfulness

Suppose that, after implementing the two approaches described in the preceding paragraph, conditional on each subset C_{sub} of the original baseline variables C, (i) X and Y remained dependent given M within all joint levels of L of the baseline variables and, indeed, that (ii) the distribution of the data (X, L, M, Y) was without any marginal or conditional independences. We now describe a novel extension of faithfulness-based causal inference that can, surprisingly, sometimes discover causal structure even when the distribution of (X, L, M, Y) is without any independences.

The novel extension is to test whether X and Y are independent conditional on both M and a subset C_{sub} of C in a weighted pseudo-population in which each subject is given either the weight $W = 1/f\left(M \mid L, X, C_{sub}\right)$ equal to the inverse of the density of M given X, L, and C_{sub} or each subject is given the stabilized weight

$$SW = f\left(M \mid C_{sub}\right)/f\left(M \mid L, X, C_{sub}\right).$$

Similar weighted pseudo-populations arise in the estimation of marginal structural models. This test can be carried out by fitting a weighted logistic regression of Y on X, M, and C_{sub} with weights given by estimates of W or of SW as in Robins *et al.*[34]. If, within strata of C_{sub}, any of the three causal DAGs in Figure 5 generated the data, then X and Y are independent conditional on M, and C_{sub} in the pseudo population. Every causal DAG in Figure 5 agrees on the following: (a) the distribution of the data (X, L, M, Y) is without any marginal or conditional independences (due to d-separation), (b) L is either an effect of X or shares a common cause with X, (c) without data on L, the effect of M on Y would be intractably confounded, (d) the direct effect null hypothesis (6) and thus the exclusion restriction holds, and (e) $r(m) = E\left[Y_{x,m} - Y_{x,m=0}\right]$ is given by $E_{pseudo}\left[Y|M = m\right] - E_{pseudo}\left[Y|M = 0\right]$, where the subscript *pseudo* denotes a pseudo-population mean in a stratum of C_{sub}.

The three DAGS in Figure 5 are the only known causal graphs for which both the pseudo-population conditional independence of X and Y given M holds and the actual distribution of (X, M, Y, L) is without any independences. We therefore believe that, under an extended faithfulness assumption[31,14], that assumes all non-trivial conditional and unconditional independences 'found' in either the population or a weighted pseudo-population are due to causal structure, one can conclude from the pseudo-population conditional independence of X and Y given M and C_{sub} that (i) X has no direct effect on Y not through M and (ii) the causal effect $r(m) = E\left[Y_{x,m} - Y_{x,m=0}\right]$ of M on Y is identified and equals $E_{pseudo}\left[Y \mid M = m\right] - E_{pseudo}\left[Y \mid M = 0\right]$.

Conclusion

The faithfulness and extended faithfulness assumptions, even if true, do not license causal conclusions in the absence of conditional independence between X and Y given M and C_{sub} either in the actual population or in a weighted pseudo-population. In realistic studies, owing to sampling variability and low power, even when the p-value of a test for conditional independence is close to 1.0, we cannot rule out weak conditional dependence. Thus, unfortunately, causal conclusions based on a faithfulness or extended faithfulness analysis must be viewed as provisional, rather than as definitive or confirmatory.

References

[1] Z. Cai, M. Kuroki, J. Pearl, and J. Tian. Bounds on Direct Effects in the Presence of Confounded Intermediate Variables. *Biometrics*, 64(3):695–701, 2008.

[2] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

[3] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[4] C. Glymour, P. Spirtes, and T.S. Richardson. On the possibility of inferring causation from association without background knowledge. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, chapter 9, pages 323–331. AAAI/MIT Press, Menlo Park, CA, 1999.

[5] C. Glymour, P. Spirtes, and T.S. Richardson. Response to rejoinder. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, chapter 11, pages 343–345. AAAI/MIT Press, Menlo Park, CA, 1999.

[6] A. N. Glynn and K. M. Quinn. Why process matters for causal inference. Unpublished manuscript, 2009.

[7] D.M. Hafeman and T.J. VanderWeele. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, ?(?):??–??, 2009.

[8] P.W. Holland. Statistics and causal inference (C/R: P961-970). *Journal of the American Statistical Association*, 81:945–960, 1986.

[9] M. G. Hudgens, P. B. Gilbert, and S. G. Self. Endpoints in vaccine trials. *Statistical Methods in Medical Research*, 13(2):89–114, 2004.

[10] K. Imai, L. Keele, and T. Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. Technical report, Department of Politics, Princeton University, 2009.

[11] Y. Jemiai, A. Rotnitzky, B. E. Shepherd, and P. B. Gilbert. Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 69(5):879–901, 2007.

[12] J.S. Kaufman, R.F. Maclehose, and S. Kaufman. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol. Perspect. Innov.*, 1(1):4, 2004.

[13] S. Kaufman, J.S. Kaufman, R.F. MacLehose, S. Greenland, and C. Poole. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine*, 24(11):1683–1702 Correction 25(18), 3228, 2005.

[14] S.L. Lauritzen, T.S. Richardson, J.M. Robins, and I. Shpitser. A markov theory for pseudo-interventions. Technical report, University of Washington, 2009.

[15] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008.

[16] J. Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 435–44, San Francisco, CA, 1995. Morgan Kaufmann.

[17] J. Pearl. *Causality.* Cambridge University Press, Cambridge, UK, 2000.

[18] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

[19] J. Pearl. Direct and indirect effects. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 411–42, San Francisco, CA, 2001. Morgan Kaufmann.

[20] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, San Mateo, CA, 1991. Morgan Kaufmann.

[21] M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(17):276–284, 2006.

[22] T. Richardson and J.M. Robins. Alternative graphical causal models and the identification of direct effects. Technical report, Department of Epidemiology, Harvard School of Public Health, 2009.

[23] J. Robins, A. Rotnitzky, and S. Vansteelandt. Discussion of *Principal stratification designs to estimate input data missing due to death* by Frangakis, C.E., Rubin D.B., An, M., MacKenzie, E. *Biometrics*, 63(3):650–653, 2007.

[24] J. M. Robins. Adjusting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.

[25] J. M. Robins. Comment on "Covariance adjustment in randomized experiments and observational studies" (Pkg: P286-327). *Statistical Science*, 17(3):309–321, 2002.

[26] J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[27] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, volume 116 of *IMA Vol. Math. Appl.*, pages 1–94. Springer, New York, 2000.

[28] J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

[29] J. M. Robins and L. Wasserman. On the impossibility of inferring causation from association without background knowledge. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, chapter 8, pages 305–321. AAAI/MIT Press, Menlo Park, CA, 1999.

[30] J. M. Robins and L. Wasserman. Rejoinder to Glymour and Spirtes. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, chapter 10, pages 333–342. AAAI/MIT Press, Menlo Park, CA, 1999.

[31] J. M. Robins and L. Wasserman. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, chapter 12, pages 349–405. AAAI/MIT Press, Menlo Park, CA, 1999.

[32] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[33] J.M. Robins. Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, Oxford, UK, 2003.

[34] J.M. Robins, M. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[35] D. B. Rubin. More powerful randomization-based *p*-values in double-blind trials with non-compliance (Pkg: P251-389). *Statistics in Medicine*, 17:371–385, 1998.

[36] D. B. Rubin. Direct and indirect causal effects via potential outcomes. *Scand. J. Statist.*, 31(2):161–170, 2004.

[37] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* Number 81 in Lecture Notes in Statistics. Springer-Verlag, 1993.

[38] T.J. VanderWeele. Bias formulas for sensitivity analysis for direct and indirect effects. Unpublished manuscript, 2009.

[39] T.J. VanderWeele and J. M. Robins. Minimal sufficient causation and directed acyclic graphs. *Ann. Statist.*, In Press:??–??, 2009.

Appendix A: Lemmas and Proofs

Lemma 1 *M-monotonicity and randomization of X i.e. Eq.* (1)*, imply*

$$M_1 \perp\!\!\!\perp Y_{00} \mid M_0 = 0 \quad \textit{if and only if} \quad X \perp\!\!\!\perp Y_{00} \mid M = 0.$$

Proof: $\mathrm{E}[Y_{00} \mid X = 0, M = 0] = E\left[Y_{00} \mid M_0 = 0\right]$ by randomization of X.

$$E\left[Y_{00} \mid M = 0, X = 1\right] = E\left[Y_{00} \mid M_1 = 0\right] = E\left[Y_{00} \mid M_1 = 0, M_0 = 0\right]$$

by randomization of X and M-monotonicity. The lemma follows directly. □

Lemma 2 *Under M-monotonicity, and randomization of X, i.e. Eq.* (1)*,*

$$\mathrm{PSDE}\,(0) \quad = \quad E\left[Y \mid X = 1, M = 0\right] - h(\mathrm{OR}_0),$$

where

$$\begin{aligned} h(\mathrm{OR}_0) &\equiv \left(1-b-\sqrt{(1-b)^2-4ac}\,\right)\Big/ (2a), \\ a &\equiv \left(1-\frac{P(M=0 \mid X=1)}{P(M=0 \mid X=0)}\right)(1-\mathrm{OR}_0)\,, \\ b &\equiv \left(\frac{P(M=0 \mid X=1)}{P(M=0 \mid X=0)} - E[Y \mid M=0, X=0]\right)(1-\mathrm{OR}_0)\,, \\ c &\equiv E[Y \mid M=0, X=0]. \end{aligned}$$

Proof: It follows directly from M-monotonicity and randomization of X that

$$\begin{aligned} E\left[Y_{1,0} \mid M_1 = M_0 = 0\right] &= E\left[Y_{1,0} \mid M_1 = 0\right] \\ &= E\left[Y_{1,0} \mid M_1 = 0, X = 1\right] \\ &= E\left[Y \mid M = 0, X = 1\right]. \end{aligned}$$

Next we observe that randomization of X implies

$$\begin{aligned} E\left[Y \mid M = 0, X = 0\right] &= E\left[Y_{0,0} \mid M = 0, X = 0\right] \\ &= E\left[Y_{0,0} \mid M_0 = 0\right] \\ &= E\left[Y_{0,0} \mid M_1 = 0, M_0 = 0\right] P(M_1 = 0 \mid M_0 = 0) \\ &\quad + E\left[Y_{0,0} \mid M_1 = 1, M_0 = 0\right] P(M_1 = 1 \mid M_0 = 0). \end{aligned} \tag{8}$$

By M-monotonicity and randomization of X we have that

$P(M_1 = 0, M_0 = 0) = P(M = 0 \mid X = 1),$ and $P(M_0 = 0) = P(M = 0 \mid X = 0),$

hence $P(M_1{=}0 \mid M_0{=}0) = P(M{=}0 \mid X{=}1)/P(M{=}0 \mid X{=}0)$. By definition

$$\mathrm{OR}_0 = \frac{E\left[Y_{0,0} \mid M_1 = 1, M_0 = 0\right](1 - E\left[Y_{0,0} \mid M_1 = 0, M_0 = 0\right])}{(1 - E\left[Y_{0,0} \mid M_1 = 1, M_0 = 0\right])E\left[Y_{0,0} \mid M_1 = 0, M_0 = 0\right]}. \tag{9}$$

Thus (8) and (9) are two equations in the two unknowns, $E[Y_{0,0}|M_1{=}1, M_0{=}0]$ and $E[Y_{0,0}|M_1{=}0, M_0{=}0]$. Solving for the latter leads directly to $h(\mathrm{OR}_0)$. □

Lemma 3 *Under M-monotonicity and randomization of X, i.e. Eq.* (1)*,* PSDE(0) =

0 *implies the following bounds:*

$$-\min\left\{P(Y{=}1 \mid M{=}0, X{=}0), \left(\frac{P(M{=}0 \mid X{=}0)}{P(M{=}0 \mid X{=}1)} - 1\right) P(Y{=}0 \mid M{=}0, X{=}0)\right\}$$
$$\leq \mathrm{RD}(0) \leq$$
$$\min\left\{P(Y{=}0 \mid M{=}0, X{=}0), \left(\frac{P(M{=}0 \mid X{=}0)}{P(M{=}0 \mid X{=}1)} - 1\right) P(Y{=}1 \mid M{=}0, X{=}0)\right\}.$$

If RD(0) *obeys these bounds then (without additional assumptions) the null hypothesis that* PSDE(0) = 0 *cannot be rejected, i.e. the bounds are tight. Here it is implicit that the hypothesis of M-monotonicity is not rejected, i.e.* $(P(M = 0 \mid X{=}0)\,/P(M{=}0 \mid X{=}1)) - 1 \geq 0$.

We provide a proof below. This result also follows directly from results of Hudgens *et al.*[9].

Proof: By M-monotonicity and randomization (1),

$$P(M{=}0 \mid X{=}1) = P(M_0{=}0, M_1{=}0) \leq P(M_0{=}0) = P(M{=}0 \mid X{=}0).$$

Further, $P(M_0{=}0, M_1{=}1) = P(M{=}0 \mid X{=}0) - P(M{=}0 \mid X{=}1)$. Now,

$$P(Y{=}1 \mid M{=}0, X{=}1) = P(Y_{10}{=}1 \mid M_0{=}0, M_1{=}0), \qquad (10)$$

since, by M-monotonicity, $\{M{=}0, X{=}1\}$ implies $\{M_0{=}0, M_1{=}0\}$. Further,

$$\begin{aligned} &P(Y{=}1 \mid M{=}0, X{=}0) \\ &= P(Y_{00}{=}1 \mid M_0{=}0, M_1{=}0)P(M_1{=}0 \mid M_0{=}0) \\ &\quad + P(Y_{00}{=}1 \mid M_0{=}0, M_1{=}1)P(M_1{=}1 \mid M_0{=}0) \\ &= P(Y_{00}{=}1 \mid M_0{=}0, M_1{=}0)P(M{=}0 \mid X{=}1)/P(M{=}0 \mid X{=}0) \\ &\quad + P(Y_{00}{=}1 \mid M_0{=}0, M_1{=}1) \times \qquad (11) \\ &\qquad (P(M{=}0 \mid X{=}0) - P(M{=}0 \mid X{=}1))\,/P(M{=}0 \mid X{=}0). \end{aligned}$$

It follows directly from (11) that we have the following bounds:

$$\max\left\{0,\ 1 - P(Y{=}0 \mid M{=}0, X{=}0)\frac{P(M{=}0 \mid X{=}0)}{P(M{=}0 \mid X{=}1)}\right\}$$
$$\leq P(Y_{00}{=}1 \mid M_0{=}0, M_1{=}0) \leq \qquad (12)$$
$$\min\left\{P(Y{=}1 \mid M{=}0, X{=}0)\frac{P(M{=}0 \mid X{=}0)}{P(M{=}0 \mid X{=}1)},\ 1\right\}.$$

If PSDE(0) = 0 then

$$\begin{aligned} P(Y_{00}\!=\!1 \mid M_0\!=\!0, M_1\!=\!0) &= P(Y_{10}\!=\!1 \mid M_0\!=\!0, M_1\!=\!0) \\ &= P(Y\!=\!1 \mid M\!=\!0, X\!=\!1). \end{aligned}$$

by (10). The bounds on RD(0) then follow by substituting $P(Y\!=\!1 \mid M\!=\!0, X\!=\!1)$ for $P(Y_{00}\!=\!1 \mid M_0\!=\!0, M_1\!=\!0)$ in (12) and then subtracting $P(Y\!=\!1 \mid M\!=\!0, X\!=\!0)$ throughout.

Tightness is a consequence of (11) together with the fact that there are no other equations linking $P(Y_{00}\!=\!1 \mid M_0\!=\!0, M_1\!=\!0)$, $P(Y_{10}\!=\!1 \mid M_0\!=\!0, M_1\!=\!0)$ and $P(Y_{00}\!=\!1 \mid M_0\!=\!0, M_1\!=\!1)$ to the law of the observed data. □

Lemma 4 *Under M-monotonicity and randomization of X and M, i.e. Eqs.* (1) *and* (3), PSDE(1) = RD(1) *and* PSDE(0) = RD(0).

Proof: We wish to show

$$E\left[Y_{11} - Y_{01} | M_1 = M_0 = 1\right] = E\left[Y | X = 1, M = 1\right] - E\left[Y | X = 0, M = 1\right].$$

First,

$$\begin{aligned} E\left[Y \mid X = 0, M = 1\right] &= E\left[Y_{01} \mid X = 0, M_0 = 1\right] \\ &= E\left[Y_{01} \mid X = 0, M_1 = M_0 = 1\right] \\ &= E\left[Y_{01} \mid M_1 = M_0 = 1\right] \end{aligned}$$

where the first equality is by consistency, the second by M-monotonicity and the last by X randomized. Next

$$\begin{aligned} E\left[Y \mid X = 1, M = 1\right] &= E\left[Y_{11} \mid X = 1, M = 1\right] \\ &= E\left[Y_{11} \mid X = 0, M = 1\right] \\ &= E\left[Y_{11} \mid X = 0, M_1 = M_0 = 1\right] \\ &= E\left[Y_{11} \mid M_1 = M_0 = 1\right] \end{aligned}$$

where the first equality is by consistency, the second by joint randomization of X and M, the third by M-monotonicity and the last by randomization of X.□

Lemma 5 *Under strong randomization of X and M, i.e. Eqs.* (1) *and* (5), $\mathrm{PSDE}(1) = \mathrm{RD}(1)$ *and* $\mathrm{PSDE}(0) = \mathrm{RD}(0)$.

Proof:

$$\begin{aligned}
&E\left[Y_{1,m} - Y_{0,m} \mid M_0 = M_1 = m\right] \\
&\quad= \quad E\left[Y_{1,m} \mid M_0 = M_1 = m, X = 1\right] - E\left[Y_{0,m} \mid M_0 = M_1 = m, X = 0\right] \\
&\quad= \quad E\left[Y_{1,m} \mid M_1 = m, X = 1\right] - E\left[Y_{0,m} \mid M_0 = m, X = 0\right] \\
&\quad= \quad E\left[Y_{1,m} \mid M = m, X = 1\right] - E\left[Y_{0,m} \mid M = m, X = 0\right] \\
&\quad= \quad \mathrm{RD}(m).
\end{aligned}$$

where the first equality follows from Eq. (1), the second follows by Eq. (5) and the third by consistency. □

Appendix B: Minimal Counterfactual (MC) Models

The MC model associated with Figure 1(b) asserts the following independence relations:

$$Y_{x,m}, M_x \perp\!\!\!\perp X,$$

for each $x \in \{0, 1\}$ and possible value for m;

$$Y_{1,m} \perp\!\!\!\perp I(M_1 = m) \mid X = 1$$

and

$$Y_{0,m} \perp\!\!\!\perp I(M_0 = m) \mid X = 0$$

where $I(M_x = m)$ is the Bernoulli indicator. In the case where M is binary the indicators in the latter two independence statements may be dropped, and the relations simplify to:

$$Y_{1,m} \perp\!\!\!\perp M_1 \mid X = 1$$

and

$$Y_{0,m} \perp\!\!\!\perp M_0 \mid X = 0.$$

In this case, the MC model is equivalent to an FFRCISTG model[32].