First version: November 14, 2011
Revised: May 24, 2012
Second Revision: June 7, 2012

TECHNICAL REPORT
R-389
June 2012

Interpretable conditions for identifying direct and indirect effects

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
(310) 825-3243 / judea@cs.ucla.edu

2nd Revision, June 7, 2012

Abstract

This paper translates the conditions necessary for the identification of natural direct and indirect effects into a transparent language, thus permitting a more informed judgment of the plausibility of these conditions. We show that the conditions usually cited in the literature are overly restricted, and can be relaxed substantially, without compromising identification. In particular, we show that natural effects can be identified by methods other than adjustment. The identification conditions can be further relaxed in parametric models with interactions, and permit us to compare the relative importance of several pathways, mediated by interdependent variables.

1 Introduction

It is well known that, unlike the assumptions needed for identifying the controlled direct effects (CDE), the identification of "natural" effects is more intricate, and requires additional assumptions (Robins and Greenland, 1992; Pearl, 2001).[1] Conditions for identifying natural direct and indirect effects are commonly derived in terms of independencies among counterfactual variables (Pearl, 2001; Robins, 2003; Petersen et al., 2006; VanderWeele and Vansteelandt, 2009; Imai et al., 2010). These independencies, often decorated with labels such as "ignorability," "conditional ignorability," or "sequential ignorability," are incomprehensible to ordinary mortals and are invoked "largely because they justify the use of available statistical methods and not because they are truly believed" (Joffe et al., 2010). Several attempts have been made recently to interpret these conditions in more conceptually meaningful way, so as to enable researchers to judge whether the necessary assumptions are scientifically plausible (Imai et al., 2011; Muthén, 2011; Valeri and VanderWeele, 2011). These interpretations avoid "ignorability" vocabulary and invoke instead such notions as

[1] The natural direct effect is defined as the effect transmitted from X and Y while keeping some intermediate variable M at whatever level it attained prior to the transition.

"no unmeasured confounders," "as if randomized," or "essentially random" which are more appealing to data analysts.

Unfortunately, these interpretations are laden with two other sources of ambiguity. First, the notion of a "confounder" is ambiguous. Some define a "confounder" (of X and Y) as a variable that affects both X and Y, some as a variable that is associated with both X and Y and still others allow for a confounder to affect X and be associated with Y. Worse yet, the expression "no unmeasured confounders" is sometimes used to exclude the very existence of such variables and sometimes to affirm the ability to neutralize them by controlling other variables. Second, the interpretations have taken "sequential ignorability" as a starting point and consequently are overly stringent – sequential ignorability is a sufficient but not necessary condition for identifying natural effects. Weaker conditions can be articulated in a transparent and unambiguous language which provide a greater identification power and a greater conceptual clarity.

A typical example of overly stringent conditions that can be found in the literature reads as follows:

> "Imai et al. (2010) showed that the sequential ignorability assumption must be satisfied in order to identify the average mediation effects. This key assumption implies that the treatment assignment is essentially random after adjusting for observed pretreatment covariates and that the assignment of mediator values is also essentially random once both observed treatment and the same set of observed pretreatment covariates are adjusted for." (Imai et al., 2011)

We shall show that milder conditions are in fact sufficient for identification; first the treatment assignment need not be random under any adjustment and, second, we need not insist on using "the same set of observe pretreatment covariates," two separate sets can sometimes accomplish what the same set does not.

Another common conception among researchers assumes that control of confounding between the treatment and the mediator can be accomplished independently of how we control the relationship between the mediator and the outcome. We will show this not be the case; adjusting for mediator-outcome confounders may confound the treatment-mediator relationship (see Fig. 11 below), therefore, adjustment for the latter should consider the covariates used in the former.

The purpose of this note is to offer a concise list of conditions that are sufficient for identifying the natural direct effect (the same holds for the indirect effect), milder than those articulated in Imai et al. (2010, 2011), and still expressed in a familiar and accessible format.

2 The Counterfactual Derivation of Natural Effects

To make this paper self-contained, Appendix A provides a formal proof of the conditions for direct-effect identification, as it appeared in (Pearl, 2001). It starts with the counterfactual definition of the natural direct effect, and then goes through three steps. First, it seeks a set of covariates W that reduces nested counterfactuals to simple counterfactuals.

Second, it reduces all counterfactuals to *do*-expressions, that is, expressions that are estimable from controlled randomized experiments. Finally it poses conditions for identifying the do-expressions from observational studies. These three steps are echoed in the informal conditions articulated in the next section. (See also Shpitser and VanderWeele (2011) and Shpitser (2012) for refinements and elaborations.)

3 Interpretable Conditions

3.1 Preliminary Notation and Nomenclature

In this section, we will try to avoid counterfactual and "ignorability" vocabulary to the maximum extent possible, and will invoke instead the notions of conditioning, confounding and independence among factors that may influence observed variables. We will say that the relationship between X and Y is "unconfounded" if the factors that influence X are independent of all factors that influence Y when X is held fixed.[2] Given a set W of covariates, we will say that "W renders a relationship unconfounded" if the relationship is unconfounded in every stratum $W = w$ of W. Such conditional unconfoundedness holds, for example, if W consists of all common causes of X and Y, but may hold for other types of covariates as well (known as "sufficient" or "admissible" (Pearl, 2009a, p. 80)). Finally, we will use the expression "W deconfounds a relationship" as a short-hand substitute for "W renders a relationship unconfounded."

We will let X stand for the treatment, M for a mediator, and Y for the outcome. We will focus our discussion on the natural direct effect, labeled NDE, though all conditions are applicable to the indirect effect as well, by virtue of the pseudo-additive decomposition of the total effect (see Pearl 2001, Eq. (23).) NDE is defined as the expected increase in Y when X is incremented by one unit and M is kept constant at what ever level it attains (in each individual) just before incrementing X.

Finally, we will assume that readers are familiar with the notion of (nonparametric) identifiability as applied to causal or counterfactual relations (see, for example (Pearl, 2009a, p. 77)). In particular, we will say that the "W-specific" causal effect of X on Y is identifiable, if the effect is consistently estimable from the observed data for every stratum level w.

3.2 Sufficient conditions for identifying natural effects

The following are two sets of interpretive conditions, marked A and B, that are sufficient for identifying natural effects, both direct and indirect. Each condition is communicated by a verbal description followed by its formal expression, where Y_{XM} stands for all factors that affect Y when X and M are held constant (at any values). Each set of conditions is followed by its graphical version, marked A_G and B_G with all graphs representing nonparametric

[2]This definition, formally written as $X \perp\!\!\!\perp Y_X$, provides a conceptually meaningful interpretation of "strong ignorability," and is given a graphical representation in Pearl (2009a, pp. 341–344). I am grateful to an anonymous reviewer for noting that the zero-bias definition given in an earlier version of this paper is insufficient – functional models must be invoked. The distinction between functional and interventional models is given in Pearl (2009a, pp. 22–38) and is further elaborated in Robins and Richardson (2010).

structural equation models (Pearl, 2009a, [Ch. 7). Set B is the stronger of the two, and represents assumptions commonly invoked in the mediation literature (Imai et al., 2010, 2011; Shpitser and VanderWeele, 2011; VanderWeele and Vansteelandt, 2009; Vansteelandt, 2012). Set A is weaker, and echoes more faithfully the derivation in Appendix A.

A (Weak conditions)

There exists a set W of measured covariates such that:

A-1 No member of W is affected by treatment.

A-2 W deconfounds the mediator-outcome relationship (holding X constant).

$$[M_X \perp\!\!\!\perp Y_{MX} \mid W]$$

A-3 The W-specific effect of the treatment on the mediator is identifiable by some means.

$$[P(m \mid do(x), w) \text{ is identifiable}]$$

A-4 The W-specific joint effect of {treatment+mediator} on the outcome is identifiable by some means.

$$[P(y \mid do(x, m), w) \text{ is identifiable}]$$

A_G (Graphical version of A)

There exists a set W of measured covariates such that:

A_G-1 No member of W is a descendant of X.

A_G-2 W blocks all backdoor paths from M to Y, disregarding the one through X.

A_G-3 The W-specific effect of X on M is identifiable using *do*-calculus.

A_G-4 The W-specific joint effect of $\{X, M\}$ on Y is identifiable using *do*-calculus.

B (stronger condition)

There exists a set W of measured covariates such that:

B-1 No member of W is affected by the treatment.

B-2 W deconfounds the treatment-{mediator, outcome} relationship.

$$[X \perp\!\!\!\perp Y_{XM}, M_X \mid W]$$

B-3 W and X deconfound the mediator-outcome relationship.

$$[M \perp\!\!\!\perp Y_M \mid X, W]$$

B_G (graphical version of B)

There exists a set W of measured covariates such that:

B_G-1 No member of W is a descendant of X.

B_G-2 W blocks all backdoor paths from X to M.

B_G-3 W and X block all backdoor paths from M to Y.

Remarks

Assumption set A differs from B on two main provisions. First, A-3 and A-4 permit the identification of causal effects by any methods whatsoever, while B-2 and B-3 require that identification be accomplished by adjustment. Second, whereas A-3 and A-4 makes no commitment to which covariates are invoked in the identification of the causal effects needed, B requires that the same set W satisfy both B-2 and B-3.

It should be noted that, whereas this paper concerns identification in observational studies, conditions A-3 and A-4 open the door to experimental studies, when such are feasible. For example, one may venture to estimate the causal effect of X on M by randomizing X or by using encouragement designs (instrumental variables) or multi-state adjustment.[3] Only the latter will be considered here. The restrictions on all such designs are the same, namely, that they be W-specific, where W is a set of attributes satisfying A-1 and A-2.

Appendix B explains why we must insist that W will be unaffected by the treatment.

4 Illustrations

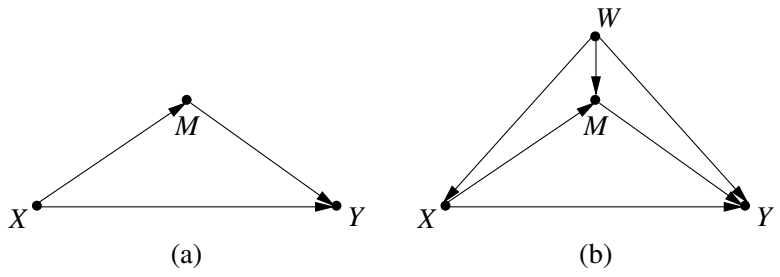

Figure 1: (a) The basic unconfounded mediation model, showing the treatment (X) mediator (M) and outcome (Y). (b) The mediator model with an added covariate (W) that confounds both the $X \rightarrow M$ and the $M \rightarrow Y$ relationships.

[3]Note that the corresponding assumption in B-3 is stronger than the zero-bias condition

$$P(y \mid do(m), x, w) = P(y \mid m, x, w)$$

which can be verified experimentally. B-3 requires that, conditioned on X and W, M be independent of the joint set $\{Y_{m_1}, Y_{m_2}, \ldots, Y_{m_k}\}$, where $\{m_1, m_2, \ldots, m_k\}$ are the values of M. Such joint independence implies zero bias, but the converse may not hold.

To illustrate and compare the conditions articulated in the previous section we start with simple models that satisfy the strong conditions of B (and B_G), and then examine how the process of identification can benefit from the relaxed conditions given in A (and A_G).

4.1 How the natural effects are identified

Figure 1(a) illustrates the classical mediation model, with no confounding; all omitted factors (not shown in the diagram) affecting X, M, and Y are assumed to be independent, so both the mediation process, $X \rightarrow M$, and the outcome process $\{X, M\} \rightarrow Y$, are said to be "unconfounded." In this model, the null set $W = \{0\}$ satisfies the conditions in B (as well as in A), and the natural direct effect takes the form[4]

$$NDE = \sum_m [E(Y \mid X = 1, M = m) - E(Y \mid X = 0, M = m)]P(M = m \mid X = 0). \quad (1)$$

Likewise, the natural indirect effect (formally defined in Appendix A) becomes[5]

$$NIE = \sum_m E(Y \mid X = 0, M = m)[P(M = m \mid X = 1) - P(M = m \mid X = 0)] \quad (2)$$

We see that both equations invoke weighted averages over the levels of M, to allow for heterogeneous populations where M and its effects on Y vary from individual to individual and cannot be equalized by policy intervention. NDE measures the portion of the total effect that would be transmitted to Y absent M's ability to detect changes in X, while NIE measures the portion transmitted absent Y's ability to detect such changes, except those transmitted through M.

Figure 1(b) illustrates a confounded mediation model in which a variable W (or a set of variables) confounds all three relationships in the model. Because W is not affected by X and is observed, adjusting for W renders all relationships unconfounded and the conditions of B (as well as A) are satisfied. Accordingly, the natural direct effect estimand is given by

$$NDE = \sum_m \sum_w [E(Y \mid X = 1, M = m, W = w) - E(Y \mid X = 0, M = m, W = w)] \\ P(M = m \mid X = 0, W = w)P(W = w) \quad (3)$$

Figure 2(a) illustrates a mediation model in which W is partitioned into three independent sets, each confounding one of the relationship in the model. It is a special case of Fig. 1(b), with $W = \{W_1, W_2, W_3\}$ which, when substituted in Eq. (3), yields:

[4]Eqs. (1) and (2) were called "The Mediation Formula" in (Pearl 2009a, p. 132; Pearl 2009b, 2011). Summations should be replaced by integration when applied to continuous variables, as in (Imai et al., 2010).

[5]The NDE and NIE are connected to each other via TE, the total effect, by the pseudo-additive relation: $TE = NDE - NIE_r$ where NIE_r is the NIE for the reversed transition, from $X = 1$ to $X = 0$. It is identified therefore whenever NDE and TE are, and all our discussions concerning the NDE should apply to NIE as well (Pearl, 2009a, p. 132; Pearl, 2009b, 2011).

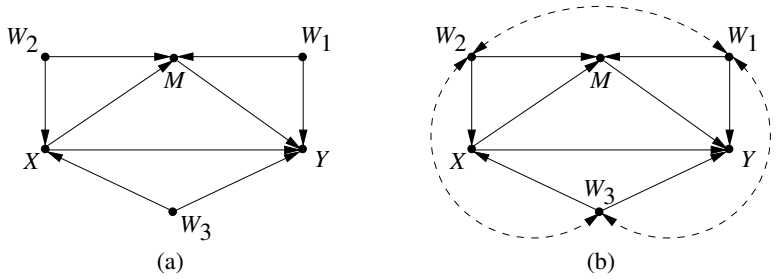

Figure 2: (a) A mediation model with three independent confounders, W_1, W_2 and W_3. (b) showing dependencies among the confounders W_1, W_2 and W_3.

$$
\begin{aligned}
NDE = \sum_{m} \sum_{w_2,w_3,w_1} & P(W_2 = w_2, W_3 = w_3, W = w_1)[E(Y \mid X = 1, M = m, W_1 = w, W_2 = w_2, W_3 = w_3)) \\
& - E(Y \mid X = 0, M = m, W_1 = w_1, W_2 = w_2, W_3 = w_3,)] \\
& P(M = m \mid X = 0, W_1 = w_1, W_2 = w_2, W_3 = w_3) \qquad (4) \\
= \sum_{m} \sum_{w_2,w_3,w_1} & P(W_2 = w_2)P(W_3 = w_3)[E(Y \mid X = 1, M = m, W_1 = w_1, W_3 = w_3)) \\
& - E(Y \mid X = 0, M = m, W_1 = w_1, W_3 = w_3)] \\
& P(M = m \mid X = 0, W_2 = w_2, W_1 = w_1)P(W_1 = w_1) \qquad (5)
\end{aligned}
$$

This decomposition is enabled by the assumption that W_1, W_2 and W_3 are mutually independent. In the more general case, as the one is in Fig. 2(b) confounders will not be independent and Eq. (3) needs to be used.

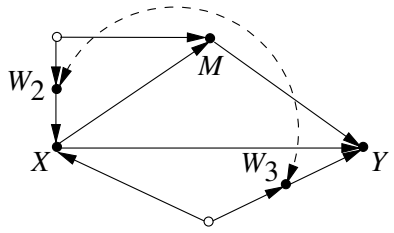

Figure 3: A mediation model with two dependent confounders, permitting the decomposition of Eq. (6). Hollow circles stand for unmeasured confounders. The model satisfies condition A but violates condition B.

4.2 Comparing identification power

We are now ready to explore the benefits of the weaker conditions expressed in A. First, note that conditions A-3 and A-4 allow for covariates outside W to assist in the identification. This results in a greater flexibility in allocating covariates for the various adjustments invoked

in expression (2). It also simplifies the process of justifying the assumptions supported by these adjustments, and leads, in turns, to a simpler overall estimand.

Specifically, in choosing covariates to deconfound the $\{X, M\} \rightarrow Y$ relationship one is free to ignore the covariates chosen to deconfound the $X \rightarrow M$ relationship, and vice versa.

The model in Fig. 3 demonstrates this flexibility. Since the mediator-to-outcome relationship is unconfounded (for fixed X), we are at liberty to choose $W = \{0\}$ to satisfy conditions A-1 and A-2. Likewise, the treatment-mediator relationship is unconfounded when we adjust for W_2, and this permits us to remove W_3 from the the factor $P(M = m \mid X = 0, W = w)P(W = w)$ of Eq. (3).[6] Finally, the $\{XM\} \rightarrow Y$ relationship is deconfounded by W_3 alone, which permits us to remove W_2 from the factors $E(Y \mid X = 1, M = m, W = w)$ and $E(Y \mid X = 0, M = m, W = w)$ of Eq. (3). The resulting estimand for NDE becomes:

$$NDE = \sum_{m} \sum_{w_2, w_3} P(W_2 = w_2, W_3 = w_3)[E(Y \mid X = 1, M = m, W_3 = w_3)] \\ - [E(Y \mid X = 0, M = m, W_3 = w_3,)]P(M = m \mid X = 0, W_2 = w_2) \qquad (6)$$

with only one of W_3 and W_2 appearing in each of the last two factors.

Note that covariates need not be pretreatment to ensure identification; B and A require merely that W be causally unaffected by the treatment. Indeed, W_3 in Fig. 3 may well be a post-treatment variable, the control of which is essential for identifying NDE. Note also that "treatment assignment" in the model of Fig. 3 is not truly "random" for it is affected by the hidden common causes of X, W_2, and W_3. It is for this reason that the term "unconfounded" is less ambiguous than "randomized."

We are now ready to compare the identification power of A versus B.

A draws its increased power from two sources:

(a) Divide and conquer – Covariates may be found capable of deconfounding the mediator and outcome processes separately but not simultaneously

(b) Multi-step adjustment – Covariates may be found capable of identifying causal effects through a nonstandard, multi-step adjustment but not through a one-step adjustment, as required by B.

(a) Divide and conquer

Fig. 3 demonstrates how the "divide and conquer" flexibility translates into an increase identification power. Here, the $X \rightarrow M$ relationship requires an adjustment for W_2, and the $X \rightarrow Y$ relationship requires an adjustment for W_3. If we make the two adjustments separately, we can identify NDE by the estimand of Eq. (6). However, if we insist on adjusting for W_2 and W_3 simultaneously, as required by assumption set B, the $X \rightarrow M$ relationship would become confounded, by opening two colliders in tandem along the path $X \rightarrow W_3 \leftrightarrow W_2 \leftarrow M$. As a result, assumption set B would deem the NDE to be unidentifiable; there is no covariates set W that simultaneously deconfounds the two relationships.

[6] This is licensed by the fact that W_2 is a sufficient deconfounder, and is valid regardless of whether W_2 and W_3 are correlated.

(b) Multi-step adjustments

Figure 4 displays a model for which the natural direct effect achieves its identifiability through multi-step adjustment (in this case using the front-door procedure), permitted by A, though not through a single-step adjustment, as demanded by B. In this model, the null set $W = \{0\}$

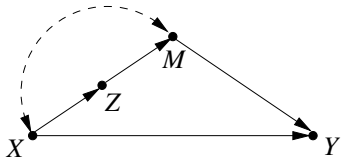

Figure 4: Measuring Z permits the identification of the effect of X on M through the front-door procedure.

satisfies conditions B-1 and B-3, but not condition B-2; there is no set of covariates that would enable us to deconfound the treatment-mediator relationship. Fortunately, condition A-3 requires only that we identify the effect of X on M by *some* method, not necessarily by rendering X random or unconfounded (or ignorable). The presence of observed variable Z permits us to identify this causal effect using the front door condition (Pearl, 1995, 2009a). The resultant NDE estimand will be:

$$NDE = \sum_m [E(Y \mid X = 1, M = m) - E(Y \mid X = 0, M = m)]P(M = m \mid do(X = 0)) \quad (7)$$

where $P(M = m \mid do(X = 0))$ is given by the front-door estimator:

$$\sum_z P(Z = z \mid X = 0, M = m) \sum_{x'} P(M = m \mid Z = z, X = x')P(X = x').$$

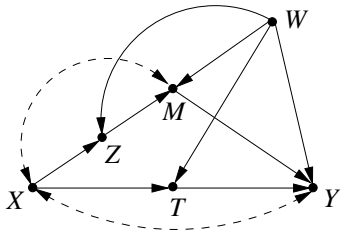

Figure 5: Measuring Z and T permits the identification of the effect of X on M and X on Y for each specific w and leads to the identification of the natural direct effect.

Figure 5 embodies the same idea in a slightly more complicated context. Here, the front-door estimator needs to be applied to both the $X \rightarrow M$ and the $X \rightarrow Y$relationships. In addition, conditioning on W is necessary, in order to satisfy condition A-2. Still, the identification of $P(M = m \mid do(x), w)$ and $E(Y \mid do(m, x), w)$ presents no special problems to students of causal inference (Shpitser and Pearl, 2006). This example demonstrates the impact of requiring A-4; had T not been observed, conditions A-1 to A-3 would have been satisfied, but not A-4.

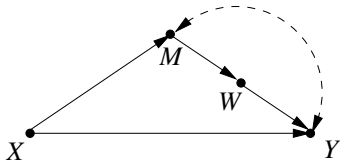

Figure 6: NDE is not identifiable because condition A-2 cannot be satisfied, though all causal effects are identifiable.

5 Coping with Treatment-dependent Confounders

Figure 6 tempts us to apply the front-door estimator to the $M \rightarrow Y$ relationship, which is confounded by unobserved common causes of M and Y (represented by the dashed arc). Unfortunately, although the causal effect of $\{X, M\}$ on Y is identifiable, condition A-2 cannot be satisfied; no covariate can be measured that deconfounds the $M \rightarrow Y$ relationship.

This is the first example we encountered in which the natural direct effect is not identifiable and the controlled direct effect is.[7] Another such example is shown in Fig. 7. Here,

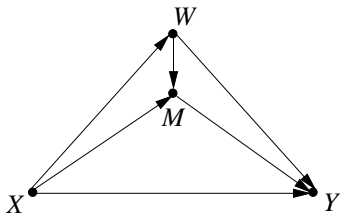

Figure 7: NDE is not identifiable because condition A-1 cannot be satisfied – W is a descendant of X.

W can serve to deconfound both the $M \rightarrow Y$ and the $X \rightarrow M$ relationships but, alas, W is a descendant of X and so, it violates condition A-1 and renders NDE non-identifiable. Figure 7 unveils a general pattern that prevents identification in any (nonparametric) model (Robins, 2003; Avin et al., 2005): Whenever a variable exists (be it measured or unmeasured) that is a descendant of X and an ancestor of both M and Y (W in our examples), NDE is not identifiable.

This restriction however does not apply to linear models, where parameter identification is all that is needed for the identification of all effects, even when a confounder W of $M \rightarrow Y$ is affected by the treatment. The same holds for other parametric models, such as linear models with interaction terms.[8]

[7] A full characterization of the conditions identifying the controlled direct effect

$$CDE(m) = E[Y \mid do(X = 1, M = m)] - E[Y \mid do(X = 0, M = m)],$$

in the presence of unmeasured confounders is given in (Shpitser and Pearl, 2008), based on the *do*-calculus.

[8] Such models have been analyzed extensively in the literature, some using purely statistical approach (MacKinnon, 2008; Kraemer et al., 2008; Jo, 2008; Preacher et al., 2007) and some applying the Mediation Formula of equations (1) and (2) (VanderWeele and Vansteelandt, 2009; Muthén, 2011; Imai et al., 2010;

To illustrate, consider the parametric version of Fig. 7:

$$y = \beta_1 m + \beta_2 x + \beta_3 xm + \beta_4 w + \epsilon_1 \tag{8}$$
$$m = \gamma_1 x + \gamma_2 w + \epsilon_2 \tag{9}$$
$$w = \alpha x + \epsilon_3 \tag{10}$$

with $\beta_3 xm$ representing an interaction term, then the basic definition of the natural effects (Appendix A) gives (for the transition from $X = 0$ to $X = 1$):

$$NDE = \beta_2 + \alpha\beta_4 \tag{11}$$
$$NIE = \beta_1(\gamma_1 + \alpha\gamma_2) \tag{12}$$
$$TE = \beta_2 + (\gamma_1 + \alpha\gamma_2)(\beta_3 + \beta_1) + \alpha\beta_4 \tag{13}$$
$$TE - NDE = (\beta_1 + \beta_3)(\gamma_1 + \alpha\gamma_2) \tag{14}$$

We see that, due to treatment-mediator interaction, $\beta_3 xm$, the portion of the effect for which mediation is *necessary* ($TE - NDE$) can differ significantly from the portion for which mediation is *sufficient* (NIE) (Pearl, 2011). The fact the W is affected by the treatment does not hinder the identification of these effects (as long as the structural parameters are identifiable), though the choice of terms for each of those effects is not trivial, and needs to be guided carefully by the formal, counterfactual definitions of NDE and NIE (Pearl, 2012).

For nonparametric models, Avin et al. (2005) derived a necessary and sufficient condition for identifying (natural) path-specific effects in graphs with no confounders. For example, suppressing the $X \to W$ or $X \to M$ processes in Fig. 7 would lead to identifiable effects, while suppressing the $W \to Y$ or $M \to Y$ processes will not.

Figure 7 can in fact be regarded as having two interacting mediators, M and W, and the results of Avin et al. (2005) highlight a fundamental difference between the two. Whereas effects mediated through W are identifiable, those mediated through M are not. For example, the natural direct and indirect effects viewing W as the mediator can be obtained directly from Eqs. (1) and (2), exchanging m with w, since the relationships $X \to W$ and $(XW) \to Y$ are unconfounded. This gives

$$NDE(W) = \sum_w [E(Y \mid X = x', W = w) - E(Y \mid X = x, W = w)]P(W = w \mid X = x)$$
$$NIE(W) = \sum_w E(Y \mid X = x, W = w)[P(W = w \mid X = x') - P(W = w \mid X = x)]$$

in which M is not invoked.

For comparison, the parametric version of Fig. (7) given in Eqs. (8)–(10) yields the following effects when W is considered the mediator:

$$NDE(W) = \beta_2 + \gamma_1\beta_1 \tag{15}$$
$$NIE(W) = \alpha(\beta_4 + \gamma_2\beta_1) \tag{16}$$
$$TE = \beta_2 + (\gamma_1 + \alpha\gamma_2)(\beta_3 + \beta_1) + \alpha\beta_4 \tag{17}$$
$$TE - NDE(W) = \alpha(\gamma_2\beta_3 + \beta_4 + \gamma_2\beta_1 + \gamma_1\beta_3) \tag{18}$$

Valeri and VanderWeele, 2011; Pearl, 2010, 2011). However, the problem of dealing with two interacting mediators (e.g., M and W in Fig. 7) has not received much attention.

Comparing Eqs. (15)–(18) to Eqs. (11)–(14) allows an investigator to assess the relative contribution of each mediator, W and M, to the overall effect of X on Y.

Figure 8 depicts the parameterized model of Eqs. (8)–(10) and compares the subgraphs carrying the effects (NIE) mediated by M and W, respectively.

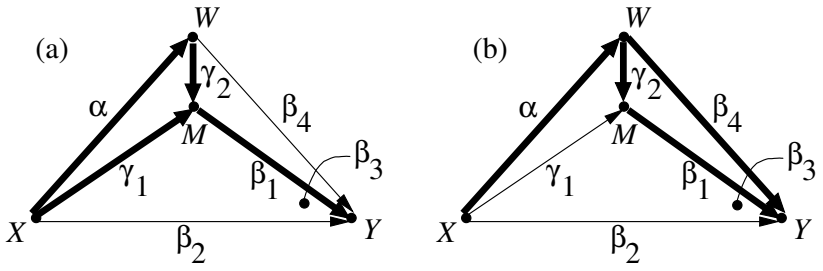

Figure 8: A parameterized version of Fig. 7, in which the heavy arrows represent (a) Paths carrying the natural indirect effect (NIE) when M is considered as the mediator. (b) Same with W considered as the mediator.

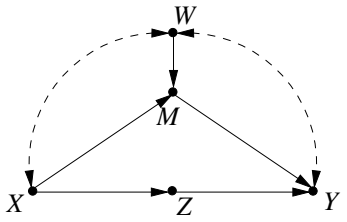

Figure 9: NDE is identified by adjusting for W and using Z to deconfound the $X \rightarrow Y$ relationship.

Figure 9, demonstrates the role that an observed covariate (Z) on the $X \rightarrow Y$ pathway can play in the identification of natural effects. In this model, conditioning on W deconfounds both the $M \rightarrow Y$ and $Y \rightarrow M$ relationships but confounds the $X \rightarrow Y$ relationship. However the W-specific joint effect of $\{X, M\}$ on Y is identifiable through observations on Z (using the front-door estimand).

Our examples thus far may create the impression that covariates situated along the path from M to Y are useless for identifying NDE. This is not the case. In Fig. 10, the mediator$\rightarrow$ outcome relationship is unconfounded (once we fix X), so, we are at liberty to choose $W = \{0\}$ to satisfy condition A-2. The treatment$\rightarrow$ mediator relationship is confounded, and requires an adjustment for T (so does the treatment-outcome relationship). However, conditioning on T will confound the $\{MX\} \rightarrow Y$ relationship (in violation of condition A-4). Here, the presence of Z comes to our help, for it permits us to estimate $P(Y \mid do(x, m), t)$ thus rendering NDE identifiable.

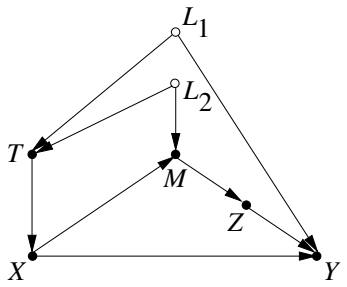

Figure 10: The confounding created by adjusting for T can be removed using measurement of Z.

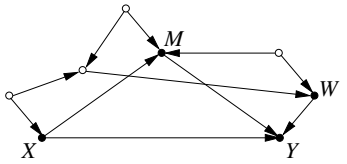

Figure 11: NDE is not identifiable because adjusting for W confounds the $X \to M$ relationship, while leaving W unadjusted renders the $M \to Y$ relationship confounded.

5.1 Is the "W-specific" provision necessary?

Assumptions A-3 and A-4 insist on identifying both the $X \to M$ and $\{X, M\} \to Y$ effects after conditioning on W. Figure 11 demonstrates that these provisions are not entailed by the others; they may produce bias if ignored. In this model, W deconfounds the $M \to Y$ and $X \to Y$ relationship, while the $X \to M$ relationship is unconditionally unconfounded. Since W must be adjusted (to deconfound the $M \to Y$ relationship), it must also be included in the deconfounding of $X \to M$. This, according to Fig. 11 is untenable (due to the collider present) and the NDE cannot therefore be identified in this model.

Fortunately, only sets that are necessary for deconfounding $M \to Y$ must participate in the deconfounding of other relationships. Fig. 3 for example shows that, since W_3 is not necessary for deconfounding $M \to Y$, it is not required to be included in the adjustment for $X \to M$. Likewise, W_2 is ignored in deconfounding $\{XM\} \to Y$. This flexibility renders NDE identifiable.

6 Conclusions

We presented a concise and interpretable list of conditions that are sufficient for the identification of natural effects and demonstrated by examples that the weaker set of conditions can lead to improved identification power. In particular, the new conditions open the door for identification methods that go beyond standard adjustment for covariates. Applying these conditions to linear models with interaction terms, we showed how path-specific effects can be estimated in models with multiple pathways and interacting mediators.

Acknowledgments

This paper has benefitted from discussions with Kosuke Imai, Salvatore Marcantonio, Bengt Muthén, and Tyler VanderWeele.

This research was supported in parts by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665 and #N00014-10-1-0933.

Appendix A

Formal Derivation of Conditions for NDE Identification (after Pearl (2001))

A.0 Notation

Throughout our analysis we will let X be the control variable (whose effect we seek to assess), and let Y be the response variable. We will let Z stand for the set of all intermediate variables between X and Y which, in the simplest case considered, would be a single variable M as in Fig. 1.[9] Most of our results will still be valid if we let Z stand for any set of such variables, in particular, the set of Y's parents excluding X.

We will use the counterfactual notation $Y_x(u)$ to denote the value that Y would attain in unit (or situation) $U = u$ under the control regime $do(X = x)$. See Pearl (2000, Ch. 7) for formal semantics of these counterfactual utterances. Many concepts associated with direct and indirect effect require comparison to a reference value of X, that is, a value relative to which we measure changes. We will designate this reference value by x^*.

A.1 Controlled Direct Effects (review)

Definition 1 *(Controlled unit-level direct-effect; qualitative)*

A variable X is said to have a controlled direct effect on variable Y in model M and situation $U = u$ if there exists a setting $Z = z$ of the other variables in the model and two values of X, x^* and x, such that

$$Y_{x^*z}(u) \neq Y_{xz}(u) \tag{19}$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ when we keep all other variables Z fixed at z. If condition (1) is satisfied for some z, we say that the transition event $X = x$ has a controlled direct-effect on Y, keeping the reference point $X = x^*$ implicit.

Clearly, confining Z to the parents of Y (excluding X) leaves the definition unaltered.

[9] The mediator is labeled Z in this appendix, to be faithful to the original derivation; it is totally interchangeable with M as used in the text.

Definition 2 *(Controlled unit-level direct-effect; quantitative)*
Given a causal model M with causal graph G, the controlled direct effect of $X = x$ on Y in unit $U = u$ and setting $Z = z$ is given by

$$CDE_z(x, x^*; Y, u) = Y_{xz}(u) - Y_{x^*z}(u) \tag{20}$$

where Z stands for all parents of Y (in G) excluding X.

Definition 3 *(Average controlled direct effect)*
Given a probabilistic causal model $\langle M, P(u)\rangle$, the controlled direct effect of event $X = x$ on Y is defined as:

$$CDE_z(x, x^*; Y) = E(Y_{xz} - Y_{x^*z}) \tag{21}$$

where the expectation is taken over u.

The distribution $P(Y_{xz} = y)$ can be estimated consistently from experimental studies in which both X and Z are randomized. In nonexperimental studies, the identification of this distribution requires that certain "no-confounding" assumptions hold true in the population tested. Graphical criteria encapsulating these assumptions are described in Pearl (2000, Sections 4.3 and 4.4).

A.2 Natural Direct Effects: Formulation

Definition 4 *(Unit-level natural direct effect; qualitative)*
An event $X = x$ is said to have a natural direct effect on variable Y in situation $U = u$ if the following inequality holds

$$Y_{x^*}(u) \neq Y_{x, Z_{x^*}(u)}(u) \tag{22}$$

In words, the value of Y under $X = x^*$ differs from its value under $X = x$ even when we keep Z at the same value ($Z_{x^*}(u)$) that Z attains under $X = x^*$.

We can easily extend this definition from events to variables by defining X as having a *natural* direct effect on Y (in model M and situation $U = u$) if there exist two values, x^* and x, that satisfy (22). Note that this definition no longer requires that we specify a value z for Z; that value is determined naturally by the model, once we specify x, x^*, and u. Note also that condition (22) is a direct translation of the court criterion of sex discrimination in hiring (Pearl, 2001) with $X = x^*$ being a male, $X = x$ a female, and $Y = 1$ a decision to hire.

If one is interested in the magnitude of the natural direct effect, one can take the difference

$$Y_{x, Z_{x^*}(u)}(u) - Y_{x^*}(u) \tag{23}$$

and designate it by the symbol $NDE(x, x^*; Y, u)$ (acronym for Natural Direct Effect). If we are further interested in assessing the average of this difference in a population of units, we have:

Definition 5 *(Average natural direct effect)*
The average natural direct effect of event $X = x$ on a response variable Y, denoted $NDE(x, x^*; Y)$, is defined as

$$NDE(x, x^*; Y) = E(Y_{x, Z_{x^*}}) - E(Y_{x^*}) \tag{24}$$

A.4 Natural Direct Effects: Identification

As noted in (Pearl, 2001), we cannot generally evaluate the average natural direct-effect from empirical data. Formally, this means that Eq. (24) is not reducible to expressions of the form

$$P(Y_x = y) \quad \text{or} \quad P(Y_{xz} = y);$$

the former governs the causal effect of X on Y (obtained by randomizing X) and the latter governs the causal effect of X and Z on Y (obtained by randomizing both X and Z).

We now present conditions under which such reduction is nevertheless feasible.

Theorem 1 *(Experimental identification)*
If there exists a set W of covariates, nondescendants of X or Z, such that

$$Y_{xz} \perp\!\!\!\perp Z_{x^*} \mid W \qquad \text{for all } z \tag{25}$$

(read: Y_{xz} is conditionally independent of Z_{x^*}, given W), then the average natural direct-effect is experimentally identifiable, and it is given by

$$\begin{aligned} NDE(x, x^*; Y) \;=\; & \sum_{w,z} [E(Y_{xz} \mid w) - E(Y_{x^*z} \mid w)] \\ & P(Z_{x^*} = z \mid w) P(w) \end{aligned} \tag{26}$$

Proof
The first term in (24) can be written

$$\begin{aligned} & E(Y_{x,Z_{x^*}} = y) \\ & \quad = \; \textstyle\sum_w \sum_z E(Y_{xz} = y \mid Z_{x^*} = z, W = w) \\ & \qquad P(Z_{x^*} = z \mid W = w] P(W = w) \end{aligned} \tag{27}$$

Using (25), we obtain:

$$\begin{aligned} & E(Y_{x,Z_{x^*}} = y) \\ & \quad = \; \textstyle\sum_w \sum_z E(Y_{xz} = y \mid W = w) \\ & \qquad P(Z_{x^*} = z \mid W = w) P(W = w) \end{aligned} \tag{28}$$

Each factor in (28) is identifiable; $E(Y_{xz} = y \mid W = w)$, by randomizing X and Z for each value of W, and $P(Z_{x^*} = z \mid W = w)$ by randomizing X for each value of W. This proves the assertion in the theorem. Substituting (28) into (24) and using the law of composition $E(Y_{x^*}) = E(Y_{x^* Z_{x^*}})$ (Pearl 2000, p. 229) gives (26), and completes the proof of Theorem 1. □

The conditional independence relation in Eq. (25) can easily be verified from the causal graph associated with the model. Using a graphical interpretation of counterfactuals (Pearl, 2000, p. 214-5), this relation reads:

$$(Y \perp\!\!\!\perp Z \mid W)_{G_{\underline{XZ}}} \tag{29}$$

In words, W d-separates Y from Z in the graph formed by deleting all (solid) arrows emanating from X and Z.

Figure 12(a) illustrates a typical graph associated with estimating the direct effect of X on Y. The identifying subgraph is shown in Fig. 12(b), and illustrates how W separates Y from Z. The separation condition in (29) is somewhat stronger than (25), since the former implies the latter for every pair of values, x and x^*, of X (see (Pearl 2000, p. 214)).

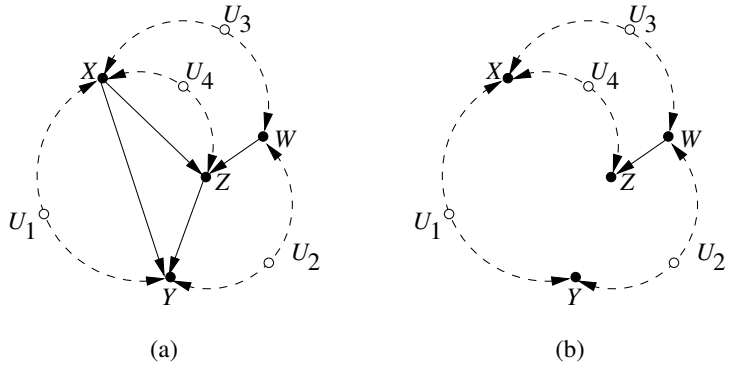

Figure 12: (a) A causal model with latent variables (U's) where the natural direct effect can be identified in experimental studies. (b) The subgraph $G_{\underline{XZ}}$ illustrating the criterion of experimental identifiability (Eq. 29): W d-separates Y from Z.

The identification of the natural direct effect from *nonexperimental* data requires stronger conditions. From Eq. (26) we see that it is sufficient to identify the conditional probabilities of two counterfactuals: $P(Y_{xz} = y \mid W = w)$ and $P(Z_{x^*} = z \mid W = w)$, where W is any set of covariates that satisfies Eq. (25) (or (29)). This yields the following criterion for identification:

Theorem 2 *(Nonexperimental identification)*
The average natural direct-effect $NDE(x, x^*; Y)$ is identifiable in nonexperimental studies if there exists a set W of covariates, nondescendants of X or Z, such that, for all values z and w we have:

(i) $Y_{xz} \perp\!\!\!\perp Z_{x^*} \mid W$

(ii) $P(Y_{xz} = y \mid W = w)$ *and* $P(Y_{x^*z} = y \mid W = w)$ *are identifiable*

(iii) $P(Z_{x^*} = z \mid W = w)$ *is identifiable*

Moreover, if conditions (i)-(iii) are satisfied, the natural direct effect is given by (26).

Appendix B

Why can't we use treatment-dependent covariates to deconfound the mediator-outcome process

Both A and B insist that no member of W be affected by the treatment, which is a requirement distinct to the identification of natural effects. For example, to identify the controlled direct effect $CDE(m)$ in Fig. 9 we can condition on $W = w$, and, using the truncated factorization formula (Pearl, 2000, p. 72), we can write

$$\begin{aligned} CDE(m) &= E[Y \mid do(X = 1, M = m)] - E[Y \mid do(X = 0, M = m)] \\ &= \sum_w E[Y \mid X = 1, M = m, W = w]P(X = 1, W = w) \\ &\quad - E[Y \mid X = 0, M = m, W = w]P(X = 0, W = w) \end{aligned}$$

The reason such conditioning does not work for the natural direct effect is that the latter is not defined in term of a population experiment (i.e., control M to level $M = m$ and change X from $X = 0$ to $X = 1$) but in terms of a hypothetical manipulation at the unit level, namely, for each individual u, freeze M at whatever level it attained for that individual, then change X from $X = 0$ to $X = 1$ and observe the change in Y).

Appendix A shows that in order to convert this unit-based operation to a population-based operation (expressible as a $do(x)$ expression) we must first find a W that deconfounds M from Y (with X fixed) and, then, conditioned on that same W, identify the counterfactual expression

$$P(M_x = m \mid W = w).$$

When W is affected by the treatment, this expression is not identifiable even when X is randomize. To see that, we recall that M_x stand for all factors affecting M when X is held fixed. These factors are none others but the omitted factors (or disturbance terms) that affect M. When we condition on W, those factors become correlated with X which renders X confounded with M.

This can also be seen from the graph, using virtual colliders. The expression $P(M_x = m \mid W = w)$ stands for the causal effect of X on M within a stratum w of W. It is identifiable using the back door criterion which demands that W not be affected by X because, as soon as W is a descendant of any intermediate variable from X to M (including M itself) a virtual collider is formed, and a new back-door path is opened by conditioning on W (Pearl, 2009a, p. 339).

Another way of seeing this, is to resort to *do*-calculus. If W is not affected by the treatment, we have $W_x = W$, and we can write

$$\begin{aligned} P(M_x = m \mid W = w) = P(M_x = m \mid W_x = w) &= \frac{P(M_x = m, W_x = w)}{P(W_x = w)} \\ &= \frac{P(M = m, W = w \mid do(X = x))}{P(W = w \mid do(X = x))} \\ &= P(M = m \mid do(X = x), W = w) \end{aligned}$$

The last expression stands for the causal effect of X on M given that $W = w$ is the post-treatment value of W. It is identifiable by the *do*-calculus, whenever the model permits such identification (Shpitser and Pearl, 2008).

It is worth mentioning at this point that treatment-dependent confounders hinder only nonparametric identification of natural effects as defined in Eq. (24). The difficulty disappears when we have a parametric representations (as in Eqs. (8)–(10)) or when we compromise on the requirement of freezing M completely at the value it attained prior to the change in treatment. For example, if in Fig. 7 we merely disable the process $X \rightarrow M$ and allow M to respond to W as we change X from $X = 0$ to $X = 1$, the resulting direct effect will be identified. This type of direct and indirect effects, which I would like to call "semi-natural effects,"[10] are defined as (using parenthetical notation):

$$SNDE = E(Y(X = 1), M(X = 0, W(X = 1)), W(X = 1)) - E(Y(X = 0))$$
$$SNIE = E(Y(X = 0), M(X = 1, W(X = 0)), W(X = 0)) - E(Y(X = 0))$$

Using the derivation leading to Eq. (26) one can show that these semi-natural effects are identifiable by:

$$SNDE = \sum_{mw} E(Y \mid X = 1, M = m, W = w)P(M = m \mid X = 0, W = w)P(W = w \mid X = 1) - E(Y \mid X = 0)$$
$$SNIE = \sum_{mw} E(Y \mid X = 0, M = m, W = w)P(M = m \mid X = 1, W = w)P(W = w \mid X = 0) - E(Y \mid X = 0)$$

Accordingly, the parametric model of Eqs. (8)–(10) would yield the following semi-natural effects:

$$SNDE = \beta_2 + \alpha(\beta_4 + \gamma_2\beta_1)$$
$$SNIE = \gamma_1\beta_1$$
$$TE = \beta_2 + (\gamma_1 + \alpha\gamma_2)(\beta_3 + \beta_1) + \alpha\beta_4$$
$$TE - SNDE = \gamma_1(\beta_1 + \beta_3) + \beta_3\alpha\gamma_2$$

Figure 13 depicts the path that supports the $SNDE$ and $SNIE$ compared with those supporting the NDE and NIE in Eqs. (11)–(14). We see that the criterion of Avin et al. (2005) is satisfied in the latter, but not the former.

References

Avin, C., Shpitser, I. and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*. Morgan-Kaufmann Publishers, Edinburgh, UK.

[10] Huber (2012) called it "partial indirect effect."

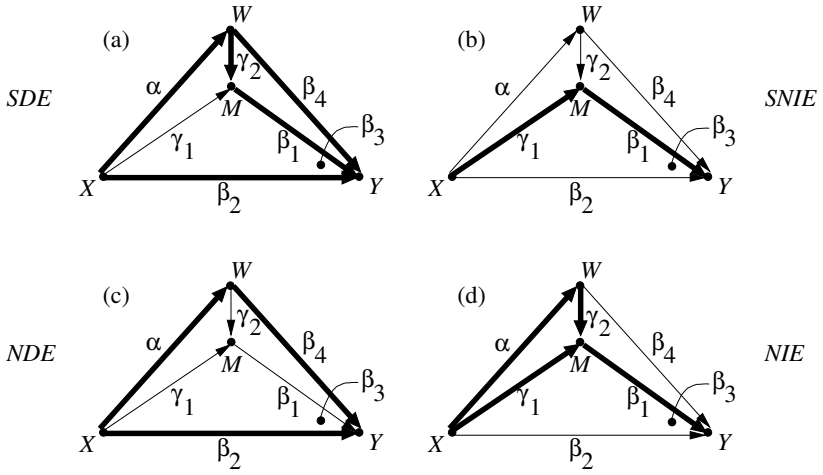

Figure 13: Subgraphs supporting the semi-natural direct and indirect effect ($SNDE$ in (a), $SNIE$ in (b)) and those supporting the natural direct and indirect effects (NDE in (c) and NIE in (d)).

Huber, M. (2012). Identifying causal mechanisms in experiments (primarily) based on inverse probability weighting. Tech. rep., University of St. Gallen, Department of Economics, Switzerland.

Imai, K., Jo, B. and Stuart, E. A. (2011). Commentary: Using potential outcomes to understand causal mediation analysis. *Multivariate Behavioral Research* **46** 842–854.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods* **13** 314–336.

Joffe, M. M., Yang, W. P. and Feldman, H. I. (2010). Selective ignorability assumptions in causal inference. *The International Journal of Biostatistics* **6**. DOI: 10.2202/1557-4679.1199.

Kraemer, H., Kiernan, M., Essex, M. and Kupfer, D. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology* **27** S101–S108.

MacKinnon, D. (2008). *An Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York.

Muthén, B. (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Tech. rep., Graduate School of Education and Information Studies, University of California, Los Angeles, CA. Submitted to *Psychological Methods*.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

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

Pearl, J. (2009a). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146. <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology* **40** 75–149.

Pearl, J. (2011). The causal mediation formula – a guide to the assessment of pathways and mechanisms. Tech. Rep. R-379, <http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, *Prevention Science*.

Pearl, J. (2012). Estimation of direct and indirect effects. In *Causality: Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). Wiley and Sons, 151–179. In print. <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>.

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Preacher, K., Rucker, D. and Hayes, A. (2007). Addressing moderated mediation hypotheses. *Multivariate Behavioral Research* **42** 185–227.

Robins, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. Green, N. Hjort and S. Richardson, eds.). Oxford University Press, Oxford, 70–81.

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

Robins, J. and Richardson, T. (2010). Alternative graphical causal models and the identification of direct effects. Tech. Rep. Working Paper no. 100, School of Public Health, Harvard University.

Shpitser, I. (2012). Counterfactual graphical models for mediation analysis via path-specific effects. Tech. rep., Harvard University, MA.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1219–1226.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* **9** 1941–1979.

Shpitser, I. and VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *The International Journal of Biostatistics* **7**. Article 16.

Valeri, L. and VanderWeele, T. (2011). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. Tech. rep., Department of Biostatistics, Harvard University. Submitted to *Psychological Methods*.

VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* **2** 457–468.

Vansteelandt, S. (2012). Estimation of direct and indirect effects. In *Causality: Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). Wiley and Sons, 126–150. In print.