The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 16

A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis

Ilya Shpitser, *Harvard University*
Tyler J. VanderWeele, *Harvard University*

Recommended Citation:
Shpitser, Ilya and VanderWeele, Tyler J. (2011) "A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 16.
DOI: 10.2202/1557-4679.1297
Available at: http://www.bepress.com/ijb/vol7/iss1/16

A Complete Graphical Criterion for the Adjustment Formula in Mediation Analysis

Ilya Shpitser and Tyler J. VanderWeele

Abstract

Various assumptions have been used in the literature to identify natural direct and indirect effects in mediation analysis. These effects are of interest because they allow for effect decomposition of a total effect into a direct and indirect effect even in the presence of interactions or non-linear models. In this paper, we consider the relation and interpretation of various identification assumptions in terms of causal diagrams interpreted as a set of non-parametric structural equations. We show that for such causal diagrams, two sets of assumptions for identification that have been described in the literature are in fact equivalent in the sense that if either set of assumptions holds for all models inducing a particular causal diagram, then the other set of assumptions will also hold for all models inducing that diagram. We moreover build on prior work concerning a complete graphical identification criterion for covariate adjustment for total effects to provide a complete graphical criterion for using covariate adjustment to identify natural direct and indirect effects. Finally, we show that this criterion is equivalent to the two sets of independence assumptions used previously for mediation analysis.

KEYWORDS: adjustment, causal diagrams, confounding, covariate adjustment, mediation, natural direct and indirect effects

Author Notes: Tyler J. VanderWeele was supported by NIH Grant HD060696.

1. Introduction

A number recent papers have developed identification and partial identification results for natural direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001; Robins, 2003; Petersen et al., 2006; Kaufman et al., 2009; Sjölander, 2009; Hafeman and VanderWeele, 2010; Imai et al., 2010ab; Robins et al. 2010; Robins and Richardson, 2010). Natural direct and indirect effects are of interest because they provide definitions of direct and indirect effects and allow for effect decomposition even in models with interaction or non-linearities. The definition of these natural direct and indirect effects employ nested counterfactual quantities that cannot be completely identified even in a doubly randomized experiment (Robins, 2003; Imai et al., 2010b).

The identification of these natural direct and indirect effects thus comes at a price, namely fairly strong assumptions that cannot be empirically confirmed. One strategy for identification is to assume that there are no interactions between the effects of the exposure and the mediatior on the outcome at the individual level (Robins and Greenland, 1992; Robins, 2003; cf. Petersen et al., 2006). Two alternative sets of identification assumptions have been proposed in the literature that allow for non-parametric identification of direct and indirect effects even in the presence of interactions. One set of assumptions, introduced by Pearl (2001), required independence of two counterfactual quantities. An alternative set of identification assumptions described in Imai et al. (2010) used other assumptions which did not require independence between two counterfactual quantities. When either of these sets of assumptions hold conditional on a single set of covariates C, natural direct and indirect effects are identified by what is sometimes referred to as the "mediation formula" (Pearl, 2010). These various assumptions raise several questions. First, we may want to know whether the two sets of identification assumptions are equivalent for some class of causal models. Second, we may want to know whether complete graphical criterion can be developed for the use of the "mediation formula" (Pearl, 2010). In this paper we answer both of these questions. Specifically, we show that for causal diagrams defined by non-parametric structural equations, the two sets of assumptions are in fact equivalent in the sense that if either set of assumptions holds for all models inducing a particular causal diagram then the other set of assumptions will also hold for all models inducing that causal diagram. Second, we give a complete graphical criterion for the use of the mediation formula for natural direct and indirect effects.

The remainder of this paper is organized as follows. In section 2 we review material on graph theory and causal diagrams. In section 3 we discuss

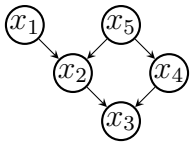


Figure 1: A directed acyclic graph.

controlled direct effects and natural direct and indirect effects, and describe two sets of assumptions given in the literature for the identification of natural effects. In section 4, we review a complete graphical identification criterion for covariate adjustment for total effects called the "adjustment criterion" (Shpitser et al., 2010) which is a generalization of Pearl's (1995) backdoor path criterion. Section 5 gives our two main results concerning the equivalence of the two sets of identification assumptions and concerning a complete graphical criterion for covariate adjustment in the identification of natural direct and indirect effects. Section 6 places the present work in the context of other forms of causal models and offers some concluding remarks. We defer the proofs and detailed technical background to the appendices.

2. Preliminaries on Graphs

We first introduce graph-theoretic terminology we need to discuss causal and probabilistic notions. A directed graph consists of a set of nodes and directed arrows connecting pairs of nodes. A path is a sequence of distinct nodes where any two adjacent nodes in the sequence are connected by an edge. A directed path from a node X to a node Y is a path where all arrows connecting nodes on the path point away from X and towards Y. If an arrow points from X to Y then X is called a parent of Y and Y a child of X. If X has a directed path to Y then X is an ancestor of Y and Y a descendant of X. By convention, X is both an ancestor and a descendant of X. A directed acyclic graph (DAG) is a directed graph where for any directed path from X to Y, Y is not a parent of X. A consecutive triple of nodes W_i, W_j, W_k on a path is called a collider if W_i and W_k are parents of W_j. Any other consecutive triple is called a non-collider. A path between two nodes X and Y is said to be blocked by a set Z if either for some non-collider on the path, the middle node is in Z, or for some collider on the path, no descendant of the middle node is in Z. For disjoint sets X, Y, Z we say X is d-separated from Y given Z if every path from a node in X to a node in Y is blocked by Z. If X is not d-separated from Y given Z, we say X is d-connected to Y given Z. See the graph in Fig. 1 for an illustration of these concepts. In this graph $X_1 \to X_2 \to X_3 \leftarrow X_4$ is a path from X_1 to X_4; $X_1 \to X_2 \to X_3$ is a directed path from X_1 to X_3; X_1

DOI: 10.2202/1557-4679.1297

is a parent of X_2, and an ancestor of X_3; $X_2 \rightarrow X_3 \leftarrow X_4$ is a collider; X_1 is d-separated from X_4 given X_5; X_1 is d-separated from X_3 given X_2 and X_4; X_1 is d-connected to X_5 given X_3.

Directed graphs play an important role in probabilistic reasoning because it is often possible to associate a graph G with a set of nodes V with a probability distribution $P(v)$ over variables V such that for any disjoint $X, Y, Z \in V$, if X is d-separated from Y given Z in G, then X is independent from Y given Z in $P(v)$. In this way, a graphical notion of path blocking represents a probabilistic notion of independence. Whenever we can associate G and $P(v)$ in this way, we say that $P(v)$ satisfies the global Markov property with respect to G.

We will represent causation using non-parameteric structural equation models (NPSEMs) (Pearl, 2009). These models consist of a set of observable variables V, and a background variable U_i for each $V_i \in V$. Each background variable U_i is assumed to vary according to some (unknown) probability distribution $P(u_i)$, while each $V_i \in V$ is given by a non-parameteric structural equation $V_i = f_i(pa_i, u_i)$, where f_i is an arbitrary function, pa_i refers to value assignments to a subset of variables in V, and u_i is a value assignment to U_i. We assume the background variables U vary according to a joint distribution $P(u)$, such that $P(u) = \prod_i P(u_i)$, in other words all background variables are jointly independent of each other. The assumption that the background variables are independent is essentially that all common causes of any two variables on the graph are also on the graph (Pearl, 2009). The distribution $P(u)$ along with the set of functions F for each $V_i \in V$ together inducing a probability distribution $P(v)$ over observable variables V in the model.

An intervention setting a variable X_i to x_i is represented in NPSEMs by replacing the function f_i for X_i by a constant-valued function evaluating to x_i. The result is a modified NPSEM with a new distribution over observables which we call an interventional distribution, and denote by $P(v|do(x_i))$.

NPSEMs are a particularly convenient formalism for reasoning about effects of interventions because the observable distribution $P(v)$ of an NPSEM satisfies the global Markov property with respect to the DAG associated with the NPSEM called a causal diagram. This association between the causal diagram and the model allows probabilistic notions such as independence, and causal notions such as confounding to be expressed in an intuitive visual way with paths for any NPSEM represented by the corresponding graph.

3. Total, Direct, and Indirect Effects

If we are interested in the effect of $do(a)$ on a subset Y of variables V, called the outcome variables, we are interested in marginal of $P(v|do(a))$, written as

$P(y|do(a))$ or $P(Y_a = y)$ where Y_a is the counterfactual value of Y that would be observed if A were set to a. This distribution represents the total effect of A on Y. The total effect may not always be the causal quantity of interest. Frequently, we may be interested in the effect of A on Y along a particular causal path. For instance, we may want to quantify the causal influence of A on Y not mediated by certain other variables in the model, i.e. the effect of A on Y if some other variables M were either held fixed, or were otherwise prevented from transmitting the influence of A on Y.

There are two formalizations of direct effects. We will need some additional notation. We let $Y_{a,m}$ denote the counterfactual value of Y that would be observed if A were set to a and M were set to m; we let M_a denote the counterfactual value of M that would be observed if A were set to a. So as to be a candidate for a mediator we restrict M to the non-ancestors of A. We also assume A is a singleton set.

Definition 1 (Controlled direct effect) Given an outcome Y, treatment value of interest a, value settings m for some other observable variables, and value settings u to all background variables, the controlled direct effect of A on Y not via M is given by $Y_{a,m}(u) - Y_{a^*,m}(u)$.

Note that the value of this effect depends on the settings of both the background variables $U = u$ and the mediating variables $M = m$. One might concieve of u as indicating a particular individual. If we wish to summarize controlled direct effect over possible values of U, we would use the average controlled direct effect (with respect to a particular setting m) $E[Y_{a,m}] - E[Y_{a^*,m}]$. The average controlled direct effect is a function of interventional distributions.

One difficulty with the controlled direct effect formulation is that it is not possible to consider indirect effects of A on Y, say the effect only along paths which include mediating variables M unless there are a set of variables that intercepts all direct paths from A to Y (VanderWeele, 2010b). An alternative which avoids this difficulty is to consider the effect of setting A to a on Y in a hypothetical situation where all the mediating variables M behaved as if a were set to a reference value a^* instead (Robins and Greenland, 1992; Pearl, 2001). Such hypothetical situations still prevent M from transmitting the influence of A on Y, yet allow a definition of an indirect effect as well, as we now show.

Definition 2 (Natural direct effect) Given an outcome Y, treatment value of interest a, a reference value a^*, and value settings u to all background variables, the natural direct effect of A on Y is given by $Y_{a,M_{a^*}}(u) - Y_{a^*,M_{a^*}}(u)$,

DOI: 10.2202/1557-4679.1297

where M_{a^*} in the subscript denotes the value of $M_{a^*}(u)$. Note that we replaced $Y_{a^*,M_{a^*}}$ by Y^*_a. These counterfactual quantities are in fact the same. The distribution $P(Y_{a^*,M_{a^*}})$ is a short-hand for writing $\sum_m P(Y_{a^*,m}, M_{a^*} = m)$ which is equal to $\sum_m P(Y_{a^*}, M_{a^*} = m) = P(Y_{a^*})$ by the generalized consistency axiom, sometimes known as composition (Pearl, 2009). See the appendix for a precise definition of generalized consistency.

As before, if we wish to summarize natural direct effect over possible values of U we would obtain the average natural direct effect $E[Y_{a,M_{a^*}}] - E[Y_{a^*}]$.

The natural indirect effect considers the behavior of the outcome Y in the situation where A is set to the reference value a^*, yet the mediating variables M vary as if A were set to a.

Definition 3 (Natural indirect effect) Given an outcome Y, treatment value of interest a, a reference value a^*, and value settings u to all background variables, the natural indirect effect of A on Y is given by $Y_{a,M_a}(u) - Y_{a,M_{a^*}}(u)$.

The average natural indirect effect, as before, is defined to be $E[Y_{a^*,M_a}] - E[Y_{a^*}]$. For the remainder of this paper, we will consider average natural direct effects, with the understanding that our results generalize in a straightforward way to indirect effects. As a terminological shorthand, we will omit "average," and simply discuss natural direct effects.

The total effect decomposes into the sum of the natural direct effect and the natural indirect effect, making these effects of particular interest in assessing the proportion of an effect mediated through an intermediate. Examples in which direct and indirect effects are of interest might include assessing the extent to which the effect of pre-eclampsia on cerebral palsy is mediated by preterm birth and whether there is a direct effect (VanderWeele and Hernández-Diaz, 2011) or assessing the extent to which certain genetic variants affect lung cancer through increased smoking or through other pathways (Chanock and Hunter, 2008).

If it is possible to implement interventions on A and M via a randomized experimental protocol, it is then possible to identify controlled direct effect, since the controlled direct effect is a function of interventional distributions. Unfortunately, in practice it is often not possible to randomize treatments of interest. Furthermore, in the case of natural direct effects, there is, in general, no experimental protocol which would identify the quantity of interest $E[Y_{a,M_{a^*}}]$ since this quantity involves A being set to one value with respect to one variable, and to another value with respect to another variable.

The next natural question is one of identification, in other words, what assumptions does one need to place on observable and counterfactual (e.g.

post-intervention) variables in order to express controlled or natural effects in terms of $P(v)$.

For controlled direct effects, a simple and complete graphical criterion is known in NPSEMs (Shpitser et al, 2006). For natural direct effects, which are a special case of counterfactual probability distributions, complete identification algorithms are also known in NPSEMs (Shpitser et al, 2008). There are three chief obstacles to applying these natural direct effect identification results in practice. First, counterfactual identification algorithms in general rely on untestable counterfactual independence assumptions. Second, they assume complete knowledge of the causal diagram, knowledge that is often unavailable in practice. Third, these algorithms may be difficult for an applied researcher to use in practice.

For the subsequent discussion, we will assume that causal knowledge, while incomplete, can be summarized as follows: there is a single treatment A, a single outcome Y, and a set of mediating variables M which lie along some causal pathways from A to Y, and a set of confounding factors C. Note that since A and Y are single variables, a singleton set containing A is denoted as $\{A\}$, and (for instance) a set containing A and all variables in C is denoted as $\{A\} \cup C$. The exact causal relationships among variables in M and C are not known. In this situation, criteria for the identification of natural direct effects exist (Pearl, 2001; Imai et al. 2010). We will use $A \perp\!\!\!\perp B|C$ to denote A is independent of B conditional on C.

Assumption set 1 (Pearl, 2001).

P1: $Y_{a,m} \perp\!\!\!\perp A|C$

P2: $Y_{a,m} \perp\!\!\!\perp M|\{A\} \cup C$

P3: $M_a \perp\!\!\!\perp A|C$

P4: $Y_{a,m} \perp\!\!\!\perp M_{a^*}|C$

Assumptions P1,P2,P3 and P4 or slight variants of them have been employed by a number of other authors (Petersen et al., 2006; van der Laan and Petersen, 2008; VanderWeele, 2009, 2010; VanderWeele and Vansteelandt, 2009).

These assumptions can all be interpreted as conditional ignorability of some outcome Y and some treatment A given some covariate set C, which is a counterfactual independence statement written as $(Y_a \perp\!\!\!\perp A|C)$. Conditional

DOI: 10.2202/1557-4679.1297

ignorability is a well known assumption which justifies adjustment for covariates C when identifying the total causal effect of A on Y from observational data. It can be interpreted as stating that there is no confounding between A and Y after we adjust for C.

Assumptions P1 and P2 can be thought of as two parts of conditional ignorability of $Y_{a,m}$ and $\{A\} \cup M$ given C, in other words $(Y_{a,m} \perp\!\!\!\perp \{A\} \cup M|C)$. Assumptions P1 and P2 are logical consequences of this conditional ignorability assumption and moreover P1 and P2 together imply $(Y_{a,m} \perp\!\!\!\perp \{A\} \cup M|C)$. The third assumption P3 is just conditional ignorability of M_a and A given C. Finally, the last assumption P4 is an independence of two counterfactual quantities. One way to interpret P4 is that it is a kind of conditional ignorability of Y_m and M given C that holds in a world where A was intervened on. However, the specific values to which A was intervened on differ between Y and M variables. It is this disagreement on values of A that makes Pearl's last assumption problematic. It is untestable. We will show later that this issue is somewhat mitigated because the last assumption is a logical implication of the other three in NPSEMs.

An alternative set of assumptions exists which avoids the difficulty of making an assumption about the independence of counterfactual quantities.

Assumption set 2 (Imai et al. 2010).

I1: $\{Y_{a,m}, M_{a^*}\} \perp\!\!\!\perp A|C$

I2: $Y_{a,m} \perp\!\!\!\perp M|\{A\} \cup C$

Assumption I2 is identical to P2, and both can be thought of as logical consequences of conditional ignorability of $Y_{a,m}$ and $\{A\} \cup M$ given C. Assumption I1 combines Pearl's first and third assumptions into a single joint statement of independence.

Hafeman and VanderWeele (2010) gave assumptions for the identification of natural direct and indirect effects that are implied by but somewhat weaker than assumption set 2; however, the results of Hafeman and VanderWeele (2010) required a binary mediator. See Robins et al. (2010), Imai et al. (2010) and Hafeman and VanderWeele (2010) for graphical and non-graphical examples in which assumption set 2 may hold without assumption set 1 holding.

Either assumption set 1 or assumption set 2 permit us to identify the natural direct effects via the following formula, which we call the adjustment formula for natural direct effects:

$$E[Y_{a,M_{a^*}}]-E[Y_{a^*}] = \sum_{c,m}\{E[Y|a,m,c]-E[Y|a^*,m,c]\}P(M=m|a^*,c)P(C=c)$$

The derivation of the adjustment formula from I1 and I2 is straightforward, we reproduce it in the appendix.

In the absence of the confounder set C this formula reduces to Pearl's mediation formula (Pearl, 2009):

$$E[Y_{a,M_{a^*}}] - E[Y_{a^*}] = \sum_{m}\{E[Y|a,m] - E[Y|a^*,m]\}(M=m|a^*)$$

Similarly, we can use the same assumptions to given the adjustment formula for natural indirect effects:

$$E[Y_{a^*,M_a}]-E[Y_{a^*}] = \sum_{c,m} E[Y|a^*,m,c]\{P(M=m|a,c)-P(M=m|a^*,c)\}P(C=c)$$

In the absence of the confounder set C this formula reduces to a version of the mediation formula for indirect effects:

$$E[Y_{a,M_{a^*}}] - E[Y_{a^*}] = \sum_{m} E[Y|a^*,m]\{P(M=m|a) - (M=m|a^*)\}$$

Using covariate adjustment with mediation formula can give rise to a particularly simple approach to the estimation of direct and indirect effects. For example, VanderWeele and Vansteelandt (2009) showed that if assumption set 2 is satisfied and if Y and M are continuous and the following regression models for Y and M are correctly specified:

$$\begin{aligned} E[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ E[M|a,c] &= \beta_0 + \beta_1 a + \beta_2' c \end{aligned}$$

then the average natural direct and indirect effects are given by

$$\begin{aligned} E[Y_{aM_{a^*}}] - E[Y_{a^*M_{a^*}}] &= \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' E[C])\}(a-a^*) \\ E[Y_{aM_a}] - E[Y_{aM_{a^*}}] &= (\theta_2\beta_1 + \theta_3\beta_1 a)(a-a^*). \end{aligned}$$

These expressions generalize the approach to mediation analysis in the social sciences of Baron and Kenny (1986) so as to allow for interactions between

DOI: 10.2202/1557-4679.1297

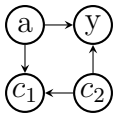


Figure 2: A graph where adjusting for $\{C_1, C_2\}$ correctly estimates the causal effect of A on Y, yet the back-door criterion fails.

the effects of the exposure and the mediator on the outcome. VanderWeele and Vansteelandt (2009) also gave expressions for the standard errors of these estimators.

4. The Adjustment Criterion for Total Effects

A problem related to identifying natural direct effects is estimating total effects, that is the effect of treatment A on outcome Y along all causal paths. Total effects are generally represented as $E[Y_a] - E[Y_{a^*}]$. A very common method for estimating causal effects assumes a set of covariates C such that adjusting for C results in an unbiased estimate of the total effect, in other words:

$$E[Y_a] - E[Y_{a^*}] = \sum_c \{E[Y|a, c] - E[Y|a^*, c]\}P(C = c)$$

This is known as the adjustment formula. Pearl gave the well known back-door criterion which permits identification of total effects by the above formula. The back-door criterion holds for C with respect to (A, Y) if C consists of non-descendants of A and d-separates all paths from A to Y which start with an arrow pointing to A (i.e. all "back-door" paths).

Unfortunately, the back-door criterion is not complete for adjustment. In other words, there exist causal diagrams where C does not satisfy the back-door criterion with respect to (A, Y), yet in all models inducing that diagram the adjustment formula above yields an unbiased estimate of the total effect.

Recently, a generalization of the back-door criterion was developed which was termed the adjustment criterion (Shpitser et al., 2010) for which we will give two further definitions.

Definition 4 (Proper Causal Path) A directed path from a node in $X \in A$ to a node in Y is called proper causal with respect to A if it does not intersect A except at X.

Definition 5 (Adjustment Criterion) The adjustment criterion holds for C with respect to (A, Y) if C blocks all paths from A to Y which are not proper causal with respect to A, and if C is not a descendant of any node on a proper causal path from A to Y (except possibly nodes in A themselves) in the graph where all arcs pointing to A are cut.

Consider the graph shown in Fig. 2. In this graph, adjusting for $\{C_1, C_2\}$ yields a valid estimate of the causal effect of A on Y, though the back-door criterion fails for $\{C_1, C_2\}$ with respect to (A, Y). Note that the set $\{C_1, C_2\}$ satisfies the adjustment criterion with respect to (A, Y) in this graph. The following results involving the adjustment criterion are known (Shpitser et al., 2010):

Theorem A. In any causal diagram G $Ç$ satisfies the adjustment criterion for (A, Y) if and only if in every NPSEM inducing G, $P(Y_a) = \sum_c P(Y|a, c)P(C = c)$.

Theorem B. In any causal diagram G, C satisfies the adjustment criterion for (A, Y) if and only if in every NPSEM inducing G, $Y_a \perp\!\!\!\perp A|C$.

In other words, the adjustment criterion characterizes both covariate adjustment and conditional ignorability in NPSEM models, which implies covariate adjustment and conditional ignorability are equivalent in NPSEMs. In fact, this equivalence holds in any causal model where conditional ignorability implies covariate adjustment, and which is a supermodel of NPSEMs. This includes almost any causal model in the literature where the generalized consistency assumption holds.

5. Adjustment Criterion for Natural Direct and Indirect Effects

Our first result is that the two sets of assumptions used for covariate adjustment for natural direct effects are in fact equivalent in NPSEMs, and are in turn equivalent to a graphical criterion which fully characterizes covariate adjustment in this setting.

Theorem 1. On any causal diagram G, assumption set 1 holds for all NPSEMs inducing G if and only if assumption set 2 holds for all NPSEMs inducing G if and only if the adjustment criterion holds for C with respect to $(\{A\} \cup M, Y)$ and for C with respect to (A, M).

A minor corollary of our result is that the set of assumptions used by Pearl is not logically minimal in NPSEMs, in the sense that one of the assumptions used is implied by the other three.

DOI: 10.2202/1557-4679.1297

Corollary 1. $(Y_{a,m} \perp\!\!\!\perp A|C)$, $(Y_{a,m} \perp\!\!\!\perp M|\{A\} \cup C)$, and $(M_a \perp\!\!\!\perp A|C)$ imply $(Y_{a,m} \perp\!\!\!\perp M_{a^*}|C)$ in NPSEMs.

Completeness of the adjustment criterion allows us to derive a complete graphical condition for the identification of natural direct and indirect effects via the adjustment formula.

Theorem 2. The adjustment formula for natural direct and indirect effects holds if and only if C satisfies the adjustment criterion relative to $(\{A\} \cup M, Y)$ and C satisfies the adjustment criterion relative to (A, M).

Note that an immediate consequence of Theorem 2 is that if an investigator believes that adjustment for C suffices to identify the joint effects of A and M on Y because of the underlying causal structure relating the variables so that

$$P(Y_{a,m}|C = c) = P(Y|A = a, M = m, C = c)$$

and that adjustment for C suffices to identify the effect of A on M because of the underlying causal structure relating the variables so that

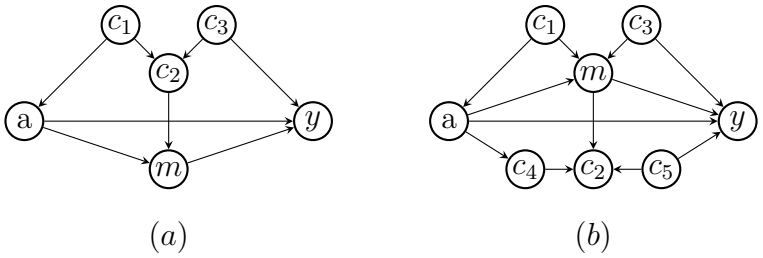
$$P(M_a|C = c) = P(M|A = a, C = c)$$

then the mediation formulas adjusting for C suffice to identify natural direct and indirect effects. We note here that these results identify the counterfactual distribution $P(Y_{a,M_{a^*}})$, rather than the expectation of Y with respect to this distribution. This means that our results apply not only to average natural direct and indirect effects, expressed in terms of expectations, but also to any function of the identified counterfactual distribution.

We now give a few positive and negative examples to illustrate our criterion. In the graph shown in Fig. 3 (a), the set $\{C_1, C_2\}$ satisfies the adjustment criterion (and the back-door criterion) with respect to (A, M), and with respect to $(\{A, M\}, Y)$. By Theorem 2, this implies that $P(Y_{a,M_{a^*}} = y)$ is identifiable in this graph, and equal to

$$\sum_{c_1,c_2,m} P(y|a, m, c_1, c_2)P(m|a^*, c_1, c_2)P(c_1, c_2)$$

At the same time, while the set $\{C_1\}$ satisfies the adjustment criterion (and the back-door criterion) with respect to (A, M), it does not satisfy either the adjustment or the back-door criterion with respect to $(\{A, M\}, Y)$. By Theorem 2, this implies $P(Y_{a,M_{a^*}} = y)$ is not identifiable by the formula



(a) (b)

Figure 3: (a) Conditions of Theorem 2 hold for the adjustment set $\{C_1, C_2\}$, with respect to (A, Y), but do not hold for the adjustment set $\{C_1\}$. (b) Conditions of Theorem 2 hold for the adjustment set $\{C_1, C_2, C_3, C_4, C_5\}$ with respect to (A, Y), but do not hold for the adjustment set $\{C_1, C_2, C_3\}$.

$$\sum_{c_1,m} P(y|a, m, c_1)P(m|a^*, c_1)P(c_1)$$

Similarly, the set $\{C_1, C_2, C_3, C_4, C_5\}$ satisfies the adjustment criterion (but not the back-door criterion) with respect to (A, M), and with respect to $(\{A, M\}, Y)$. This implies by Theorem 2 that $P(Y_{a,M_{a^*}} = y)$ is identified by the formula

$$\sum_{c_1,c_2,c_3,c_4,c_5,m} P(y|a, m, c_1, c_2, c_3, c_4, c_5)P(m|a^*, c_1, c_2, c_3, c_4, c_5)P(c_1, c_2, c_3, c_4, c_5)$$

On the other hand, the set $\{C_1, C_2, C_3\}$ fails to satisfy either the adjustment criterion or the back-door criterion with respect to both (A, M) and $(\{A, M\}, Y)$. Thus, the effect $P(Y_{a,M_{a^*}} = y)$ is not identified by the formula

$$\sum_{c_1,c_2,c_3,m} P(y|a, m, c_1, c_2, c_3)P(m|a^*, c_1, c_2, c_3)P(c_1, c_2, c_3)$$

6. Discussion

In this paper we have shown that two sets of identification assumptions for natural direct and indirect effects (Pearl, 2001; Imai et al., 2010) are in fact equivalent for NPSEMs and that, for NPSEMs, they are moreover equivalent to a simple graphical criterion that we have called "the adjustment criterion for the mediation formula." We have moreover shown that in the context of

DOI: 10.2202/1557-4679.1297

NPSEMs this graphical criterion is in fact complete for the use of the mediation formula; that is, if the graphical criterion is not satisfied for a particular causal directed acyclic graph then there is some model consistent with the diagram for which the mediation formula does not equal the natural direct and indirect effects.

We have derived our results within the context of causal diagrams that correspond to NPSEMs. We close by making a few remarks concerning our results when one steps out of the NPSEM framework. First, outside this context, the various identification assumptions need not be equivalent. Indeed, Imai, Kelle and Yamamoto point out that while assumptions I1, and I2 imply assumptions P1,P2,P3, and P4, the converse is not necessarily true. There are a number of examples in the literature showing this (Imai et al., 2010; Hafeman and VanderWeele, 2010; Robins et al., 2010). Nevertheless, we believe that NPSEMs constitute a very broad and general class of data generating mechanisms and that most of the examples in which one set of assumptions holds and the other fails are typically quite contrived. A second point is that even outside of the NPSEM framework, the completeness of our criterion will still be applicable to any alternative graphical model (Robins and Richardson, 2010) that is a supermodel of NPSEMs such as the "minimal counterfactual models" considered by Robins and Richardson (2010). Finally, we note that although our adjustment criterion is sound and complete for the use of the mediation formula on NPSEMs, it does not entail a complete identification criterion on NPSEMs more generally for natural direct and indirect effects. It is complete only for the use of covariate adjustment in the identification of natural direct and indirect effects. Although there are examples in which the adjustment criterion fails but natural direct and indirect effects are still identified, they are not identified by covariate adjustment and the use of the mediation formula. We believe that in practice there will be few, if any, settings in which a researcher will have sufficient structural knowledge to determine that the adjustment criterion for the mediation formula fails and also has sufficient structural knowledge to determine that natural direct and indirect effects are otherwise identified.

Appendix A (Technical Preliminaries)

We give background material and definitions needed for a formal proof of our results.

For a given NPSEM, its causal disagram is a directed acyclic graph with a vertex for every variable in V, and a directed arrow from V_i to V_j if $V_i \in Pa_j$. It is known that for any NPSEM with distribution $P(v)$ over observables and

a causal diagram G, $P(v)$ satisfies the global Markov property with respect to G (Verma and Pearl, 1988; Geiger et al., 1990; Lauritzen et al., 1990; Pearl, 2009). Effects of an intervention $do(v_i^*)$ in an NPSEM with a causal diagram G can be represented graphically by removing all incoming arrows to V_i, to obtain a mutilated graph $G_{\overline{v_i}}$. For any value assignment x to a set of variables $X \subseteq V$, it is known that $P(v|do(x))$ satisfies the global Markov property with respect to $G_{\overline{x}}$.

Assuming an NPSEM with a causal diagram that is a DAG amounts to assuming whenever two variables are on the graph, all their common causes are also on the graph. For this to be satisfied the causal diagram may include nodes which are unmeasured common causes. To facilitate arguments in the appendix, we assume there is an underlying causal diagram that is a DAG, that corresponds to some distribution P, but we only observe a marginal of this distribution over V, the observable variables. Such marginals may not satisfy the global Markov property with respect to any DAG. However, they may satisfy the global Markov property with respect to a mixed graph. To make our proofs as general as possible, we consider mixed graphs containing two kinds of edges, directed and bidirected. For a given DAG G containing nodes V partitioned into sets O of observed nodes and L of latent nodes, we construct the latent projection $G(O)$ of G as follows. $G(O)$ contains a vertex for every nodes in O. Furthermore, if there is a d-connected path from X to Y (given the empty set) in G where all intermediate nodes are in L, with the first arrow pointing away from X and the last arrow pointing towards Y, we add a directed arrow from X to Y in $G(O)$. If the first arrow points towards X and the last arrow points towards Y, we add a bidirected arrow from X to Y in $G(O)$. The notion of d-separation generalizes in a natural way to latent projections as m-separation (Richardson et al, 2002). A distribution P satisfies the global Markov property with respect to a mixed graph G if for any disjoint sets X, Y, Z if X is m-separated from Y given Z in G, then X is independent of Y given Z. For an NPSEM with a graph G if only a subset $O \subseteq V$ is observed, then $P(O)$ satisfies the global Markov property with respect to a latent projection $G(O)$. Representing an intervention $do(x)$ by cutting incoming arrows to X generalizes in a straightforward way to latent projections, where incoming bidirected arcs to X are cut as well. In subsequent proofs we will assume NPSEMs with a subset O of observable nodes, represented by a latent projection $G(O)$.

To derive Theorems 1 and 2, we also need to consider properties of counterfactual independence. A widely known set of properties of conditional independence are the so called graphoid axioms (Dawid 1979; Pearl 1988).

DOI: 10.2202/1557-4679.1297

- $(X \perp\!\!\!\perp Y|Z) \Leftrightarrow (Y \perp\!\!\!\perp X|Z)$
- $(X \perp\!\!\!\perp Y \cup W|Z) \Rightarrow (X \perp\!\!\!\perp Y|Z)$
- $(X \perp\!\!\!\perp Y \cup W|Z) \Rightarrow (X \perp\!\!\!\perp Y|W \cup Z)$
- $(X \perp\!\!\!\perp Y|Z) \wedge (X \perp\!\!\!\perp W|Y \cup Z) \Rightarrow (X \perp\!\!\!\perp Y \cup W|Z)$

These axioms hold in arbitrary probability distributions, and in particular in counterfactual distributions.

The next axiom we will need is the axiom of generalized consistency, sometimes referred to as composition (Pearl, 2010). In NPSEMs, this axiom states that if we observe $W_x(u) = w$, then for any Y, $Y_x(u) = Y_{x,w}(u)$.

Finally, we will need to use the axiom of compositionality, which states that if $(X \perp\!\!\!\perp Y|Z)$ and $(W \perp\!\!\!\perp Y|Z)$, then $(X \cup W \perp\!\!\!\perp Y|Z)$. Compositionality does not hold in arbitrary probability distributions. Nevertheless, compositionality does hold in distributions which satisfy a path-wise global Markov property (such as d-separation) with respect to some graph for those independences which are implied by that Markov property. This is because such Markov properties are based on paths, and statements about paths between sets decompose into statements about paths between elements of these sets. For our purposes we are interested in applying compositionality to independence statements in counterfactual distributions, and in NPSEMs there exists a graph which displays counterfactual independences via d-separation. This graph is known as a counterfactual graph (Shpitser et al, 2007).

For a given counterfactual distribution $P(\gamma) = P(Y^1_{a^1}, ..., Y^k_{a^k})$ derived from an NPSEM inducing a causal diagram G, the counterfactual graph G_γ is obtained by considering a mutilated graph $G_{\overline{a^i}}$ for each $Y^i_{a^i}$ in γ, and having these graphs share the background variables, representing the generalized consistency axiom. For a detailed construction, see (Shpitser et al, 2008). The distribution $P(An(Y^1_{a^1})_{G_\gamma}, ..., An(Y^k_{a^k})_{G_\gamma})$ satisfies the global Markov property with respect to G_γ, and thus any conditional independences due to d-separation in G_γ will satisfy the compositionality axiom.

Appendix B (Proofs)

We first show that assumptions I1 and I2 imply the adjustment formula for natural direct effects.

$$
\begin{aligned}
&E[Y_{a,M_{a^*}}] = \\
&\sum_m E[Y_{a,m}|M_{a^*} = m]P(M_{a^*} = m) = \\
&\sum_{c,m} E[Y_{a,m}|M_{a^*} = m, C = c]P(M_{a^*} = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y_{a,m}|M_{a^*} = m, C = c]P(M = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y_{a,m}|M = m, A = a^*, C = c]P(M = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y_{a,m}|A = a^*, C = c]P(M = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y_{a,m}|A = a, C = c]P(M = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y_{a,m}|M = m, A = a, C = c]P(M = m|A = a^*, C = c)P(C = c) = \\
&\sum_{c,m} E[Y|M = m, A = a, C = c]P(M = m|A = a^*, C = c)P(C = c)
\end{aligned}
$$

The first equality is by definition, and the second is by case analysis. The third is implied by the generalized consistency axiom, the fourth is implied by I1, the fifth is implied by I2, the sixth is implied by I1 and the graphoid axioms, the seventh by I2, and finally, the eighth by the generalized consistency axiom.

Theorem 1. On any causal diagram G, Assumption 1 holds for all NPSEMs inducing G if and only if Assumption 2 holds for all NPSEMs inducing G if and only if the adjustment criterion holds for C with respect to $(A \cup M, Y)$ and for C with respect to (A, M).

Proof. We first show Assumption 1 implies Assumption 2. Note that $(Y_{a,m} \perp\!\!\!\perp M|\{A\} \cup C)$ is both I2 and P2. Assumptions P1 and P3 imply I1 by compositionality. Thus P1,P2,P3 imply I1,I2. We now show the converse: again I2 and P2 are equivalent; also I1 implies P1 and P3. It remains to show I1 and I2 imply P4. In fact, I2 implies $(Y_{a,m} \perp\!\!\!\perp M_{a^*}|A = a^*, C)$ by the generalized consistency axiom. But together with I1 and the graphoid axioms, this implies $(Y_{a,m} \perp\!\!\!\perp \{M_{a^*}, A = a^*\}|C)$. But this implies $(Y_{a,m} \perp\!\!\!\perp M_{a^*}|C)$ which is P4.

Note from this argument it thus follow that P1,P2,P3 imply P4 for NPSEMs. Thus P1,P2,P3 and I1,I2 are equivalent.

DOI: 10.2202/1557-4679.1297

By results in Shpitser et al. (2010), the adjustment criterion holds for Z with respect to (X, Y) in a NPSEM if and only if $(Y_x \perp\!\!\!\perp X|Z)$ holds. This means we must show the equivalence of the two assumptions and the following two independences: $(Y_{a,m} \perp\!\!\!\perp A, M|C)$ and $(M_a \perp\!\!\!\perp A|C)$. Note that $(M_a \perp\!\!\!\perp A|C)$ is P3, while P1 and P2 are equivalent to $(Y_{a,m} \perp\!\!\!\perp A, M|C)$ by the graphoid axioms. We saw above that P1,P2,P3 and I1,I2 and P1,P2,P3,P4 are all equivalent for NPSEMs and this completes the proof.

Corollary 1 $(Y_{a,m} \perp\!\!\!\perp A|C)$, $(Y_{a,m} \perp\!\!\!\perp M|A, C)$, and $(M_a \perp\!\!\!\perp A|C)$ imply $(Y_{a,m} \perp\!\!\!\perp M_{a^*}|C)$ in NPSEMs.

Proof. This is immediate from the previous argument.

To prove completeness for Theorem 2, we will need some utility lemmas.

Lemma A. Let M be partitioned into M^1 and M^2 where M^1 is a subset of non-descendants of A. Then $P(Y_{a,M_{a^*}}) = P(Y_{a,M^2_{a^*}})$.

Proof. We have $P(Y_{a,M_{a^*}}) = \sum_m P(Y_{a,m}, M_{a^*} = m) = \sum_m P(Y_{a,m}, M^2_{a^*} = m^2, M^1_a = m) = \sum_m P(Y_{a,m^2}, M^2_{a^*} = m^2, M^1_a = m) = \sum_{m^2} P(Y_{a,m^2}, M^2_{a^*} = m^2) = P(Y_{a,M^2_{a^*}})$. Here the first identity is by rules of probability, the second is by assumption on M^1, and the third is by generalized consistency. The last identity is by definition.

Lemma B. Let M be partitioned into M^1 and M^2 where M^1 is a subset of non-ancestors of Y. Then $P(Y_{a,M_{a^*}}) = P(Y_{a,M^2_{a^*}})$.

Proof. We have $P(Y_{a,M_{a^*}}) = \sum_m P(Y_{a,m}, M_{a^*} = m) = \sum_m P(Y_{a,m^1,m^2}, M^2_{a^*} = m^2, M^1_a = m) = \sum_m P(Y_{a,m^2}, M^2_{a^*} = m^2, M^1_a = m) = \sum_m P(Y_{a,m^2}, M^2_{a^*} = m^2) = P(Y_{a,M^2_{a^*}})$. Here the first two identities are by rules of probability, the third is by assumption on M^1, and the last by definition.

Lemma C Assume A, Y, M are such that $P(Y_{a,m}) = P(Y_m)$ in every model inducing G. Then $P(Y_{a,M_{a^*}}) = P(Y)$.

Proof. We have $P(Y_{a,M_{a^*}}) = \sum_m P(Y_{a,m}, M_{a^*} = m) = \sum_m P(Y_m, M_{a^*} = m) = \sum_m P(Y, M_{a^*} = m) = P(Y)$.

The first identity is by definition, the second by assumption, the third by generalized consistency, and the last by rules of probability.

Lemmas A and B together imply that we may restrict ourselves to a mediator set M which lies in $An(Y) \cap De(A)$ without loss of generality. Lemma C implies that we can restrict ourselves to the (non-trivial) case of mediators

which leave some causal paths from A to Y open. In all subsequent results, we will assume M which lies in $An(Y) \cap De(A)$ and leaves some causal paths from A to Y open. We call such sets M standard mediating sets for (A, Y).

Lemma 1. Let M be a standard mediating set with respect to (A, Y). Then if the adjustment criterion holds for C with respect to (A, Y), then the adjustment criterion holds for C with respect to (A, M).

Proof. If there exists W on a proper causal path from A to M such that $C \cap De(W) \neq \emptyset$, then W must also lie on a proper causal path from A to Y. If C opens a non-causal path π from A to M, then adjoining a directed path from M to Y to π results in a non-causal path π^* from A to Y. Note that M must have a directed path to Y by assumption, and C cannot intersect this directed path since that would violate the adjustment criterion for C with respect to (A, Y). This implies C opens π^*, which leads to a contradiction.

Lemma 2. Let M be a standard mediating set with respect to (A, Y). If the adjustment criterion holds for C with respect to (A, Y), then C blocks all non-causal paths from A to Y, and C does not contain descendants of nodes $W \notin A \cup M$ which lie on a proper causal path from $A \cup M$ to Y.

Proof. The first claim follows by definition of the adjustment criterion. If the second claim is false, then C cannot satisfy the adjustment criterion with respect to (A, Y), since M must lie in $De(A) \cap An(Y)$.

Lemma 3. Let M be a standard mediating set with respect to (A, Y). Assume the adjustment criterion does not hold for C with respect to (A, Y). Then either the adjustment criterion does not hold for C with respect to (A, M) or the adjustment criterion does not hold for C with respect to $(\{A\} \cup M, Y)$.

Proof. If C opens a non-causal path from A to Y, then the adjustment criterion does not hold for C with respect to $(\{A\} \cup M, Y)$. If C is a descendant of a node $W \neq A$ on a proper causal path π from A to Y, then either π intersects M or not. If it does not, then the adjustment criterion does not hold for C with respect to $(\{A\} \cup M, Y)$. If it does, then there is either a node M' in M between A and W on π or not. If so, then C is a descendant of M' and thus C violates the adjustment criterion for (A, M). If not, then W is on a path from A to M which again violates the adjustment criterion for C with respect to (A, M).

Lemma 4. Fix a causal diagram G. Let $\gamma[x_1, ..., x_k]$ be a counterfactual probability which is a function of $x_1, ..., x_k$ which are subscripted values of variables $X_1, ..., X_k$ in a class of NPSEMs inducing G. Let $x'_1, ..., x'_k$ be value assignments to $X_1, ..., X_k$ different from $x_1, ..., x_k$. Assume $\gamma[x_1, ..., x_k]$ is not

DOI: 10.2202/1557-4679.1297

identifiable (in a class of NPSEMs inducing G) by a functional $\phi[x_1, ..., x_k]$ (derived from the appropriate joint distribution P of the models in the NPSEM class). In other words, there exists an NPSEM inducing G where $\gamma[x_1, ..., x_k] \neq \phi[x_1, ..., x_k]$.

Then $\gamma[x'_1, ..., x'_k]$ is not identifiable by $\phi[x'_1, ..., x'_k]$ in the same class of NPSEMs.

Proof. If $\gamma[x'_1, ..., x'_k]$ is identifiable by $\phi[x'_1, ..., x'_k]$, then variable replacement yields a contradiction.

Lemma 5. Let M be an NPSEM inducing a graph G where some nodes are deterministic functions of their parents. If there are probability distributions P_M, Q_M over observable counterfactual variables in M such that $P_M \neq Q_M$, then there exists an NPSEM M^* inducing G where no observable node is a deterministic function of its parents, such that $P_{M^*} \neq Q_{M^*}$.

Proof. Let D be the set of deterministic nodes in M. Fix an arbitrarily small $\epsilon > 0$. To construct M^*, we copy all functions in M, except for every node $X_i \in D$, we associate a new binary unobserved parent U_i independent of other unobserved variables with the function $F_i^{M^*}$ determining $X_i^{M^*}$ behaving as F_i in M with respect to arguments other than $U_i^{M^*}$ if $U_i^{M^*} = 1$, and as some function other than F_i with respect to arguments other than U_i^M if $U_i^{M^*} = 0$. For each $U_i^{M^*}$, we let $P(U_i^{M^*} = 1)$ be large enough so the probability of M^* behaving as M is greater than $1 - \epsilon$.

Thus, since models M and M^* have identical behavior with probability greater than $1 - \epsilon$, for all values v of the distributions P_M, P_{M^*}, $|P_M(v) - P_{M^*}(v)| \leq \epsilon$. Similarly, for Q_{M^*} and Q_M. It suffices to set epsilon to be less than $|P_M(v) - Q_M(v)|/2$ for some set of values v of P_M to get our conclusion.

We are now ready to prove Theorem 2.

Theorem 2. Let M be a standard mediating set with respect to (A, Y). The adjustment formula for natural direct and indirect effects holds if and only if C satisfies the adjustment criterion relative to $(\{A, M\}, Y)$ and C satisfies the adjustment criterion relative to (A, M).

Proof. We first prove soundness. If C satisfies the adjustment criterion relative to to $(\{A, M\}, Y)$ and C satisfies the adjustment criterion relative to (A, M), then the adjustment formula natural effects holds by Theorem 1 and results of Imai and Pearl.

We now prove completeness. Assume C does not satisfy the adjustment criterion for (A, Y). Then by Lemma 3, either C does not satisfy the adjustment criterion for $(\{A\} \cup M, Y)$, or C does not satisfy the adjustment criterion for (A, M).

Let $\gamma[a, a^*] = P(Y_{a,M_{a^*}} = y) = \sum_m P(Y_{a,m} = y, M_{a^*} = m)$, and $\phi[a, a^*] = \sum_{m,c} P(Y = y|c, m, a)P(m|c, a^*)P(c)$. Now by Theorem A, $\gamma[a, a]$ is not identifiable by $\phi[a, a]$, which implies our conclusion by Lemma 4.

Assume C satisfies the adjustment criterion for (A, Y). By Lemmas 1 and 2, the only way C can fail to satisfy the adjustment criterion for either $(A \cup M, Y)$ or (A, M) is if C opens a non-causal path π from $M' \in M$ to Y. Furthermore, π must be back-door for M and the arrow in π adjancent to Y must point to Y. If the former condition is not true, then a directed path from A to M' joined with π would result in a non-causal path from A to Y open by C. If the latter condition is not true, then there is an element in C which is a descendant of Y, which would mean the adjustment criterion does not hold for C with respect to (A, Y).

To prove our result, it suffices to give a counterexample model for this case such that the natural effect is not equal to the adjustment formula. We are free to choose among models not faithful to G, in other words models which contain independence restrictions not encoded by G via a path-separation criterion. Without loss of generality, we restrict ourselves to DAGs. In cases where the model is a latent projection of a DAG, we can always recover a DAG by replacing each bidirected arc connecting a node pair with a new node parent of that node pair.

Fix a single $M' \in M$ with a path π_1 from M' to Y, with the arrow in π_1 adjacent to M' pointing to M', and the arrow in π_1 adjacent to Y pointing to Y. Path π_1 is d-connected given some minimal subset C' of C. By assumption, there is a directed path π_2 from A to M'. Since M is a standard mediating set for (A, Y) by assumption, we may assume there is a directed path π_3 from A to Y.

See Fig. 4 for a schematic representation of this case. Note that our proof will work for an arbitrary graph embedding a subgraph with A, M', C', Y corresponding to this schematic representation (e.g. A has a directed path to Y and to M', and there is a path from M' to Y open by C with arrows adjacent to M' and Y pointing to M' and Y). This is because we can always consider a non-faithful model where all nodes not equal to A, M', Y and not on the paths π_1, π_2, π_3 are jointly independent of all other nodes in the model (for instance, each node could correspond to a Bernoulli random trial).

By properties of d-separation it follows that the path π_1 decomposes into a set of segments $\tau_1, ..., \tau_n$ where each segment τ_i is a directed path, the first such segment points to M', the last such segment points to Y, and if there are two intermediate segments, they point to a node ancestral of an element in C (since π_1 is d-connected given C'). Note that for odd i, τ_i and τ_{i+1} share the topmost node, we will denote it as $E_{(i+1)/2}$. Note that by definition of π_1,

DOI: 10.2202/1557-4679.1297

the topmost node, we will denote it as $E_{(i+1)/2}$. Note that by definition of π_1, $n \geq 2$ and is even, and so there are $n/2$ such topmost nodes.

We now give a partially deterministic parameterization of the family of models shown in Fig. 4. We will show this parameterization gives a counterexample showing our claim. A fully stochastic parameterization will then follow by Lemma 5.

All variables will be binary. Variables $A, E_1, ..., E_{n/2-1}$ are Bernoulli with $p = 0.5$. $E_{n/2}$ is Bernoulli with $p = \epsilon$. All other variables X_i are set to $(\sum Pa(X)) \pmod 2$, in other words their values are equal to the sum of values of their parents modulo 2. We now give a series of lemmas about this parameterization culminating in showing $P(Y_{a,M_{a^*}} = y) \neq \sum_{c,m} P(Y = y|m, c, a)P(m|c, a^*)P(c)$.

Lemma 6. In our model family, $C \perp\!\!\!\perp A$, and $P(Y_{a,M_{a^*}} = y) = P(Y_a = y) = P(y|a)$.

Proof. The first claim is a consequence of d-separation in our family of graphs. To see that the second identity is true, note that $P(Y_{a,M_{a^*}} = y) = \sum_m P(Y_{a,m} = y, M_{a^*} = m) = \sum_m P(Y_a = y, M_{a^*} = m) = P(Y_a = y)$ follows because M has no directed paths to Y. $P(Y_a = y) = P(y|a)$ follows because A and Y are not confounding in our family of graphs.

Lemma 7. Assume that $C' = \{C_1, ..., C_k\}$ (a possibly empty set) in our model family. Then the following identities hold:

(1) $P(a, c', m') = \frac{1}{2^{k+1}}(1 - \epsilon)$ if $(m' + \sum_{c_i \in c'} c_i) \pmod 2 = a$

(2) $P(a, c', m') = \frac{1}{2^{k+1}}\epsilon$ if $(m' + \sum_{c_i \in c'} c_i) \pmod 2 \neq a$

(3) $P(Y = 0|c', m', A = 0) = 1$ if $(m' + y + \sum_{c_i \in c'} c_i) \pmod 2 = a$

(4) $P(Y = 0|c', m', A = 0) = 0$ if $(m' + y + \sum_{c_i \in c'} c_i) \pmod 2 \neq a$

Proof.

First, we note that $n/2 \geq k + 1$. This is because every path segment τ_i must have a unique descendant in $C' \cup \{M'\}$, and every element in C' is a descendant of at least two path segments τ_i, τ_{i+1} in π_1.

Next, we note that $(m' + \sum_{c_i \in c} c_i) \pmod 2 = (a + e_k) \pmod 2$. This is because $\{M'\} \cup C'$ can be viewed of taking the sum $\pmod 2$ of $A, E_1, ..., E_{n/2}$, except the values of $E_1, ..., E_{n/2-1}$ are counted twice.

Note that $P(a, c', m') = \sum_{e_{n/2}} P(c', m'|e_{n/2}, a)P(a)P(e_{n/2})$. If $A = 0$ and $E_{n/2} = 0$, $P(c', m'|e_{n/2}, a) > 0$ only if $(m + \sum_{c_i \in c'} c_i) = 0$. Since $n/2 \geq k + 1$, and since $E_1, ..., E_{n/2-1}$ are Bernoulli with $p = 0.5$, $P(c', m'|e_{n/2}, a)$ is uni-

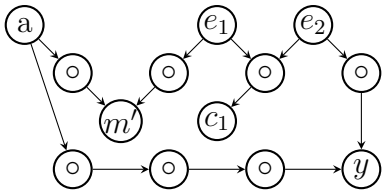


Figure 4: Schematic representation of the case involving confounding of mediator and outcome.

form with respect to c', m' for values of c', m' which have positive probability, in other words: $P(c', m'|e_{n/2}, a) = 1/2^k$ if above constraint on c', m' holds. $P(a) = 0.5$ and $P(E_{n/2} = 0) = 1 - \epsilon$, which implies (1) and (2).

Similarly, $(y + m' + \sum_{c_i \in c'} c_i) = 0 \pmod 2$ in our model. This is because $\{Y, M'\} \cup C'$ can be viewed as taking the sum $\pmod 2$ of $A, E_1, ..., E_{n/2}$, and every value is counted twice. This implies that $P(y = 0|m', c') = 1$ if and only if $(m' + \sum_{c_i \in c'} c_i) = 0$. This implies (3). and (4).

Lemma 8. In our model family, $P(Y = 0|A = 0) = 1 - \epsilon$, while $\sum_{c,m'} P(y = 0|c, m', a = 0)P(m'|c, a = 1)P(c) = \epsilon$.

Proof. Y is a function of A and $E_{n/2}$ in our model family. If A is known to equal 0, Y will equal A if and only if $E_{n/2}$ assumes value 0 which happens with probability $1 - \epsilon$.

Next, note that $P(Y = 0|c', m', A = 0)$ is only defined if $(m' + \sum_{c_i \in c'} c_i) \pmod 2 = 0$, and is then equal to 1 by Lemma 7. Finally, also by Lemma 7, $P(c', m'|A = 1) = \frac{1}{2^k}\epsilon$ if $(m' + \sum_{c_i \in c'} c_i) \pmod 2 = 0$.

Since there are k elements in C', there are 2^{k+1} terms in the sum $\sum_{c',m'} P(Y = 0|c', m', A = 0)P(c', m'|A = 1)$. In our model, only 2^k terms do not go to 0. This implies the sum evaluates to ϵ, which is our conclusion.

References

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 357-363.

Baron, R.M. and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173-1182.

DOI: 10.2202/1557-4679.1297

Chanock, S.J. and Hunter D.J. (2008) When the smoke clears. *Nature*, **452(3)**, 537-538.

Dawid A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society,,B* **41**, 1-31.

Hafeman, D.M. and VanderWeele, T.J., Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, in press.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, in press. Web address: http://imai.princeton.edu/research/files/mediation.pdf

Kaufman, S., Kaufman, J.S. and MacLehose, R.F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*, in press.

Pearl, J. (1995). Casual diagrams for empirical research (with discussion). *Biometrika*, **82**, 669-710.

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

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

Pearl, J. (2010). The mediation formula. UCLA Technical Report.

Petersen, M.L., Sinisi, S.E., and van der Laan, M.J. (2006). Estimation of direct causal effects. *Epidemiology*, **17**, 276-84.

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

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

Robins, J.M., Richardson, T.S. and Spirtes, P. (2010). On identification and inference for direct effects. *Epidemiology*, in press.

Robins, J.M., Ricahrdson, T.S. (2010). Alternative causal graphsical models and the identification of direct effects. Working paper.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In Uncertainty in Artificial Intelligence, volume 22.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, 352–359.

Shpitser, I. and Pearl, J. (2008). Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, **9**, 1941-1979.

Shpitser, I., VanderWeele, T.J. and Robins, J.M. (2010). On the validity of covariate adjustment for estimating causal effects. Uncertainty and Artificial Intelligence, in press.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, **28**, 558-571.

van der Laan, M.J. and Petersen, M.L. (2008). Direct effect models. *International Journal of Biostatistics*, **4**, Article 23.

VanderWeele, T.J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, **20**, 18-26.

VanderWeele, T.J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface - Special Issue on Mental Health and Social Behavioral Science*, 2, 457-468.

VanderWeele, T.J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, **21**, 540-551.

VanderWeele, T.J. (2010b). Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian Journal of Statistics*, in press.

VanderWeele, T.J and Hernández-Diaz, S. (2011). Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatric and Perinatal Epidemiology*, in press.

DOI: 10.2202/1557-4679.1297