Biometrics 59, 1–??
December 2006

DOI: 10.1111/j.1541-0420.2005.00454.x

Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions

Stijn Vansteelandt

Department of Applied Mathematics and Computer Sciences, Ghent University,
Krijgslaan 281 S9, 9000 Ghent, Belgium
**email:* stijn.vansteelandt@ugent.be

and

Tyler J. VanderWeele

Departments of Epidemiology and Biostatistics, Harvard School of Public Health,
677 Huntington Avenue, MA 02115 Boston, U.S.A.
**email:* tvanderw@hsph.harvard.edu

Summary: We define natural direct and indirect effects on the exposed. We show that these allow for effect decomposition under weaker identification conditions than population natural direct and indirect effects. When no confounders of the mediator-outcome association are affected by the exposure, identification is possible under essentially the same conditions as for controlled direct effects. Otherwise, identification is still possible with additional knowledge on a non-identifiable selection-bias function which measures the dependence of the mediator effect on the observed exposure within confounder levels, and which evaluates to zero in a large class of realistic data-generating mechanisms.

We argue that natural direct and indirect effects on the exposed are of intrinsic interest in various applications. We moreover show that they coincide with the corresponding population natural direct and indirect effects when the exposure is randomly assigned. In such settings, our results are thus also of relevance for assessing population natural direct and indirect effects in the presence of exposure-induced mediator-outcome confounding, which existing methodology has not been able to address.

Key words: Causal inference; Direct effect; Indirect effect; Mediation; Pathway; Time-varying confounding.

1. Introduction

For many decades, scientists in diverse scientific fields - most notably, epidemiology, psychology and sociology - have been occupied with questions as to whether an exposure affects an outcome through pathways other than those involving a given mediator / intermediate variable. The answer to such questions is of interest because it brings insight into the mechanisms that explain the effect of exposure on outcome (VanderWeele, 2009) and because the presence of intermediate variables may sometimes complicate the interpretation of the exposure effect (Joffe et al., 2001; Rosenblum et al., 2009). Mediation analyses are used for this purpose. They attempt to separate so-called 'indirect effects', which designate that part of an exposure effect which arises indirectly by affecting a (given) set of intermediate variables, from the remaining 'direct effect'.

Direct effects are traditionally connected with the conditional association between outcome and exposure, given the mediator(s); the indirect effect is typically obtained through a combination of the exposure's effect on the mediator and the mediator's effect on the outcome (Baron and Kenny, 1986; MacKinnon, 2008). For instance, when the associations between exposure A and mediator M and outcome Y can be modeled as

$$\begin{aligned} E(Y|A,M) &= \beta_0 + \beta_a A + \beta_m M \\ E(M|A) &= \alpha_0 + \alpha_a A, \end{aligned}$$

then β_a is commonly interpreted as a direct effect and $\beta_m \alpha_a$ as an indirect effect (Baron and Kenny, 1986). It is well known from the causal inference literature that these interpretations are often not justified - even when the exposure A is randomly assigned - as a result of confounding of the mediator-outcome association (Cole and Hernán, 2002); adjustment for such confounding becomes non-standard when confounders L of the mediator-outcome association are themselves affected by the exposure (see Figure 1, right) (Robins, 1999; VanderWeele, 2009; Vansteelandt, 2009b). Furthermore, decomposition of a total effect into a direct and indirect effect becomes subtle when certain nonlinear associations exist between mediator and outcome (Robins and Greenland, 1992; Pearl, 2001).

[Figure 1 about here.]

Robins and Greenland (1992) introduced model-free definitions of direct and indirect effect, which add up to the total exposure effect. Their formalism of so-called pure or natural direct effects makes use of composite counterfactuals such as $Y(a, M(a^*))$, which denotes the counterfactual outcome that would have been observed if the exposure A were set to a and the mediator M to the value $M(a^*)$ that it would have taken at some reference exposure level a^*. Because such composite counterfactuals are unobservable when $a \neq a^*$, strong assumptions are needed for identification. The development of Robins and Greenland (1992) precludes the existence of exposure effect modification by the mediator on the additive scale (at the individual level). In the development of Pearl (2001), identification is achieved by precluding the possibility of exposure-induced or 'intermediate' confounding of the mediator-outcome association. As we argue in Section 2, this places severe restrictions on the range of realistic applications that can be addressed.

In Section 3, we propose definitions of natural direct and indirect effects on the exposed, which add up to the total effect on the exposed. We show that these natural direct and indirect effects on the exposed can be identified under weaker assumptions than the corresponding population effects. In particular, they can be identified when all confounders of the exposure-outcome association and of the mediator-outcome association have been measured, and in addition, a specific, non-identifiable selection-bias function is known. This selection-bias function evaluates to zero in the absence of exposure-induced confounding, as well as under a large class of data-generating mechanisms that allow for exposure-induced confounding. We moreover show that these natural direct and indirect effects on the exposed equal the corresponding population effects when the exposure is randomly assigned. Our results thus also entail identifiability for natural direct and indirect population effects in the presence of exposure-induced mediator-outcome confounding. The prior absence of such methodology has been one of the difficulties with the causal inference literature on mediation.

These natural direct and indirect effects on the exposed are thus theoretically appealing from the standpoint of identification. They are moreover of intrinsic interest insofar as often, when overall causal effects are analyzed, it is the effect of treatment on the treated, rather than the average treatment effect for the entire population, that is in view. The effect of treatment on the treated is of relevance in evaluating what the actual effect of treatment is amongst those who took it. In an observational study, those who did not take treatment may not have done so because of some knowledge that the treatment effect was likely not to be advantageous to them. In these settings, it is of more policy relevance to evaluate what the actual effect was amongst those who in fact thought the treatment was sufficiently beneficial to make use of it. In other settings, the group who did not take treatment may include persons that are especially difficult to induce to take treatment or for whom that would be undesirable, and the question of what would happen if everyone in a population took treatment may therefore be of less policy relevance if it simply is not possible or undesirable to induce everyone to take treatment. For instance, Vansteelandt et al. (2009) argue that the effect of hospital-acquired infection on mortality is primarily relevant in those who acquired it because it is of interest to prevent infection for that subgroup, but it is not of interest to induce infection in those who remained infection-free. Treatment effects on the treated may also be of interest because of the greater accuracy with which they can sometimes be estimated. This is for instance the case when the support of the confounder distribution in the treated is strictly contained within the support of the confounder distribution in the untreated. In such settings, the data tend to be more informative about the treatment effect in the treated than in the total population (see e.g. Kurth et al. (2006)).

2. Natural direct and indirect effects

We briefly review definitions of natural direct effects and discuss limitations of the current developments which formed the motivation for this work. The (population) natural or pure direct effect (Robins and Greenland, 1992; Pearl, 2001) is defined as the expected contrast $E\{Y(a, M(a^*)) - Y(a^*, M(a^*))\}$. As a leading example to illustrate this, consider the public health question addressed in Section 5 on the direct and indirect effects of adequate ($A = 1$) or inadequate ($A = 0$) prenatal care on the risk of preterm birth (Y) other than through pre-eclampsia (M). Here, $E\{Y(1, M(0)) - Y(0, M(0))\}$ expresses the effect of adequate prenatal care on the risk of preterm birth as it would have been observed if the occurrence of pre-eclampsia were as in the absence of adequate prenatal care; alternatively, one may consider controlling the mediator at $M(1)$, rather than $M(0)$. Intuitively, the natural direct effect thus appears to capture what would be realized if the exposure was administered, but its effect on the mediator were somehow blocked; see Robins (2003) and Didelez et al. (2006) for subtleties surrounding this more intuitive interpretation.

Under the composition assumption (VanderWeele and Vansteelandt, 2009) that for each a, $Y(a, M(a)) = Y(a)$ with probability 1, the difference between the total causal effect and a natural direct effect measures an indirect effect:

$$E\{Y(a) - Y(a^*)\} - E\{Y(a, M(a^*)) - Y(a^*)\} = E\{Y(a, M(a)) - Y(a, M(a^*))\},$$

which we term the (population) natural indirect effect. In the example, $E\{Y(1, M(1)) - Y(1, M(0))\}$ would express the change in preterm birth risk under adequate care if women's pre-eclampsia status changed to what it would be without adequate prenatal care.

Identification of natural direct and indirect effects requires various assumptions besides the consistency assumptions - which we shall make throughout - that $Y(a, m) = Y$ with probability 1 amongst those with $A = a$ and $M = m$, and that $M(a) = M$ with probability 1 amongst those with $A = a$. First, it requires assumptions sufficient to identify the effects of A and M on the outcome (Pearl, 2001). VanderWeele and Vansteelandt (2009) formalize this by assuming that data are available on a set of covariates C which is sufficient to control for confounding of the effects of A and M on the outcome, in the sense that

$$Y(a,m) \perp\!\!\!\perp A|C, \tag{1}$$

and

$$Y(a,m) \perp\!\!\!\perp M|A,C, \tag{2}$$

for all a, m; here, (2) should be read as saying that for each (a, m) and each possible realization (a^*, c) of (A, C), $Y(a, m)$ is independent of M amongst subjects with $A = a^*$ and $C = c$. Second, their reliance on counterfactuals $M(a)$ demands additional assumptions sufficient to identify the effect of A on M (Pearl, 2001). VanderWeele and Vansteelandt (2009) formalize this by assuming that the same set of covariates C is also sufficient to control for confounding of the effect of A on M, in the sense that

$$M(a) \perp\!\!\!\perp A|C, \tag{3}$$

for all a. Finally, their reliance on unobservable composite counterfactuals, such as $Y(a, M(a^*))$ for $a \neq a^*$, necessitates additional identification conditions. Pearl (2001) assumes that

$$Y(a,m) \perp\!\!\!\perp M(a^*)|C, \tag{4}$$

for all a, m, which rules out the possibility of exposure-induced mediator-outcome confounding. This is an important limitation because it is often likely to believe that some prognostic factors of the mediator may themselves be affected by the exposure, especially when the mediator - and hence some of its prognostic factors - comes much later in time than the exposure (Robins, 1999); see VanderWeele and Vansteelandt (2009) for exceptions. For instance, smoking confounds the association between pre-eclampsia and preterm birth because it reduces the likelihood of pre-eclampsia and increases the likelihood of preterm birth. In addition, smoking may itself be affected by adequate prenatal care.

A number of authors have considered alternative identification conditions. Imai et al. (2010) obtain identification under assumption (2) together with the joint independence assumption $\{Y(a,m), M(a^*)\} \perp\!\!\!\perp A|C$ for all a, a^*, m, which also requires the absence of intermediate confounders. Robins and Greenland (1992) and Petersen et al. (2006) allow for intermediate confounding (i.e., exposure-induced mediator-outcome confounding) by working instead under specific no-interaction assumptions. Robins and Greenland (1992) assume the absence of exposure-mediator interactions at the individual level in the sense that $Y(a,m) - Y(0,m)$ is a random variable not depending on m. In that case, the natural direct effect equals the so-called controlled direct effect $E\{Y(a,m) - Y(a^*,m)|C\}$ for arbitrary m (Robins and Greenland, 1992), which can be identified in the presence of (measured) intermediate confounding (Robins, 1999). However, the no-interaction assumption of Robins and Greenland (1992) is strong and unlikely to hold in practice (Petersen et al., 2006). For instance, in the example, the individual effect of adequate prenatal care may potentially be higher with pre-eclampsia than without, because of the closer monitoring these women with pre-eclampsia may require. Petersen et al. (2006) assume instead that $E\{Y(a,m) - Y(a^*,m)|M(a^*) = m, C\} = E\{Y(a,m) - Y(a^*,m)|C\}$ for all a, m. This assumption is more difficult to comprehend and will be violated when the individual (controlled) direct effect of exposure on outcome interacts with some of the intermediate confounders L (in view of the association between L and $M(a^*)$). For instance, in the example it effectively presupposes that the direct effect of adequate prenatal care is the same regardless of a woman's natural pre-eclampsia status. This assumption may well be violated because the direct effect of adequate prenatal care may be higher for smoking women (in view of its potential effect on smoking), and these women are likely to have a reduced (natural) risk of pre-eclampsia.

In the next section, we will attempt to overcome the aforementioned limitations by focussing on natural direct and indirect effects on the exposed.

3. Natural direct and indirect effects on the exposed

We define the conditional natural direct effect on the exposed to be:

$$E\left\{Y - Y(a^*, M)|A, C\right\}, \tag{5}$$

and the marginal natural direct effect on the exposed to be:

$$E\left\{Y - Y(a^*, M)|A\right\}. \tag{6}$$

These express, within each exposure stratum (and possibly also within strata of baseline confounders), how much the average outcome would change if the exposure were set to a^*, but the mediator were held fixed at its *observed* level.

The above definitions allow for variation in the mediator level between subjects (unlike controlled direct effects) and enable decomposition of the total effect (on the exposed) into a direct and indirect effect (on the exposed), as follows

$$E\left\{Y - Y(a^*)|A, C\right\} = E\left\{Y - Y(a^*, M)|A, C\right\} + E\left\{Y(a^*, M) - Y(a^*)|A, C\right\}.$$

Here, the term

$$E\left\{Y(a^*, M) - Y(a^*)|A, C\right\} \tag{7}$$

can be interpreted as an indirect effect as it evaluates how much the outcome would change on average if the exposure's effect acted only through modifying the mediator. Under the aforementioned composition assumption, which we shall make throughout, we obtain that:

$$\begin{aligned}
E\left\{Y - Y(a^*)|A = a, C\right\} &= E\left\{Y(a) - Y(a^*)|A = a, C\right\}, \\
E\left\{Y - Y(a^*, M)|A = a, C\right\} &= E\left\{Y(a, M(a)) - Y(a^*, M(a))|A = a, C\right\}, \\
E\left\{Y(a^*, M) - Y(a^*)|A = a, C\right\} &= E\left\{Y(a^*, M(a)) - Y(a^*, M(a^*))|A = a, C\right\}
\end{aligned}$$

where the first expression is the total effect on those exposed to $A = a$ and the second and third expression are natural direct and indirect effects on the exposed, respectively (in the terminology of Robins and Greenland (1992), they correspond with a total direct effect and pure indirect effect). For dichotomous exposure A taking values 0 for the unexposed and 1 for the exposed, (5) with $a^* = 0$ thus equals $E\{Y(1, M(1)) - Y(0, M(1))|A = 1, C\}$ in the exposed and equals zero in the unexposed. The definitions thus suggests using $M(a)$ as a natural reference level for subjects with exposure a, which may be of interest in itself when the choice of reference levels ($M(a)$ or $M(a^*)$) seems a priori difficult to justify. Further, (7) with $a^* = 0$ equals the natural indirect effect on the exposed, $E\{Y(0, M(1)) - Y(0, M(0))|A = 1, C\}$, and equals zero in the unexposed. This motivates the choice of nomenclature for (5) and (7) as the natural direct and indirect effect on the exposed, respectively.

The following theorem shows that (conditional) natural direct and indirect effects on the exposed equal the corresponding population natural direct and indirect effects when

$$Y(a^*, M(a)) \perp\!\!\!\perp A|C, \tag{8}$$

for arbitrary a and a^*. This assumption is stronger than (1) (e.g., it could be violated if the association between A and M were confounded by unmeasured factors). It is satisfied under the causal diagrams of Figure 1 and holds in particular when the exposure is randomly assigned, conditional on C.

Theorem 1: *Under assumption (8):*

$$\begin{aligned} E\{Y - Y(a^*, M)|A = a, C\} &= E\{Y(a, M(a)) - Y(a^*, M(a))|C\} \\ E\{Y(a^*, M) - Y(a^*)|A = a, C\} &= E\{Y(a^*, M(a)) - Y(a^*, M(a^*))|C\}. \end{aligned}$$

Proof. We only give a proof of the first equality (the second is analogous):

$$\begin{aligned} E\{Y - Y(a^*, M)|A = a, C\} &= E\{Y(a, M(a)) - Y(a^*, M(a))|A = a, C\} \\ &= E\{Y(a, M(a)) - Y(a^*, M(a))|C\}. \quad \square \end{aligned}$$

This result is important as it implies that the identification results in Section 4 are also relevant for assessing population natural direct and indirect effects. The same cannot necessarily be concluded for marginal natural direct and indirect effects on the exposed because the covariate distribution in the exposed may still differ from the population distribution.

4. Identification

4.1 *Identification in the absence of intermediate confounding*

Theorem 2, whose proof is in the Web Appendix, shows that in the absence of intermediate confounding, natural direct and indirect effects on the exposed can be identified under essentially the same identification conditions, (1) and (2), as for controlled direct effects.

THEOREM 2: *Suppose that the no unmeasured confounder assumptions (1) and (2) hold for given a^*, then*

$$E\left\{Y(a^*,M)|A,C\right\} \;=\; \int E\left(Y|M=m,A=a^*,C\right)f(M=m|A,C)dm \tag{9}$$

and

$$\begin{aligned} E\left\{Y(a^*,M)|A=a\right\} &= \int E\left(Y|M=m,A=a^*,C=c\right)f(M=m,C=c|A=a)dmdc \qquad (10)\\ &= E\left\{Y\frac{I(A=a^*)}{f(A=a^*|C)}\frac{f(M=m|A=a,C)}{f(M=m|A=a^*,C)}\frac{f(A=a|C)}{f(A=a)}\right\}. \qquad (11)\end{aligned}$$

Note from Theorem 2 that the assumptions required for the identification of natural direct and indirect effects on the exposed are slightly stronger than for identification of controlled direct effects, but weaker than for identification of the corresponding population effects. Identification of controlled direct effects requires assumption (2) only for a equalling the observed exposure level A, in addition to assumption (1). Identification of population natural direct and indirect effects requires additional ignorability assumptions such as (3) with respect to the association between A and M. In contrast, we do not have to appeal to counterfactuals $M(a)$ to define natural direct and indirect effects in the exposed, so that expression (9) can be used when the association between exposure and mediator is confounded by unmeasured factors. While expression (9) for $A = a$ coincides with Pearl's expression for $E\left\{Y(a^*,M(a))\right\}$ (Pearl, 2001), the latter is only valid when the additional assumptions (3) and (4) both hold.

Expressions (10) and (11) suggest two estimation strategies for marginal natural direct and indirect effects on the exposed. The first amounts to fitting a regression model for the outcome conditional on mediator, exposure and confounders, next evaluating the fitted values at $A = a^*$ and at the observed mediator and confounder values, and finally taking their average within the subgroup with $A = a$. The second amounts to first fitting regression models for the mediator conditional on the exposure and confounders, and for the exposure conditional on the confounders. The fitted values from these models can subsequently be used to construct a weight

$$\frac{I(A = a^*)}{f(A = a^*|C)} \frac{f(M = m|A = a, C)}{f(M = m|A = a^*, C)} \frac{f(A = a|C)}{f(A = a)},$$

for each subject, on the basis of which $E\{Y(a^*, M)|A = a\}$ can be estimated as the corresponding weighted average of the outcome in the total sample.

4.2 *Identification in the presence of intermediate confounding*

Implicit behind (1) and (2) is the assumption that the set of covariates which is sufficient to adjust for confounding of the exposure-outcome association is also sufficient to adjust for confounding of the mediator-outcome association. Because it thereby precludes the possibility of intermediate confounding, we will now relax assumption (2) to the weaker assumption

$$Y(a, m) \perp\!\!\!\perp M|A, L, \tag{12}$$

for all a, m, where L is a set of covariates which must include C, but may additionally contain components some of which are affected by the exposure. This assumption is satisfied under the causal diagram of Figure 1 (right). The following identities

$$\begin{aligned} E\{Y(a^*, M)|A, C\} &= \int E\{Y(a^*, m)|M = m, A, C\} f(M = m|A, C)dm \\ &= \int E\{Y(a^*, m)|L, A\} f(L, M = m|A, C)dmdL, \end{aligned} \tag{13}$$

where we rely on assumption (12) in the second equality, make clear that this relaxation complicates identification of the natural direct effect on the exposed. In particular, whilst

$E\{Y(a^*, m)|A, C\}$ can be identified as

$$\begin{aligned} E\{Y(a^*, m)|A = a^*, C\} &= \int E\{Y(a^*, m)|L, A = a^*\} f(L|A = a^*, C)dL \\ &= \int E(Y|M = m, L, A = a^*) f(L|A = a^*, C)dL, \qquad (14) \end{aligned}$$

under assumptions (1) and (12) (Robins, 1999), neither $E\{Y(a^*, m)|M = m, A, C\}$ nor $E\{Y(a^*, m)|L, A\}$ can be identified because conditioning on M or L may render $Y(a^*, m)$ and A dependent as a result of collider-stratification. Theorem 3, whose proof is given in the Web Appendix, states our main result: it shows that progress can be made upon quantifying that degree of dependence in terms of a selection-bias function.

Theorem 3: *(i) Suppose that the no unmeasured confounder assumptions (1) and (12) hold. Suppose furthermore that the following selection-bias function is known*

$$q_a(A, m, L) = E\{Y(a, m) - Y(a, 0)|A, L\} - E\{Y(a, m) - Y(a, 0)|A = a, L\}, \qquad (15)$$

which measures to what extent the mediator effect at controlled levels a of the exposure varies over differently exposed subgroups conditional on L. Then

$$\begin{aligned} E\{Y(a^*, M)|A, C\} &= \int\int E(Y|M = m, L, A = a^*)f(L|A = a^*, C)dLf(M = m|A, C)dm \\ &+ \int\int \{E(Y|M = m, L, A = a^*) + q_{a^*}(A, m, L)\} \{f(L|M = m, A, C) - f(L|A, C)\} dL \\ &\times f(M = m|A, C)dm. \qquad (16) \end{aligned}$$

(ii) The choice of $q_a(A, m, L)$ imposes no restrictions on the observed data law in the sense for each choice of $q_a(A, m, L)$, one can find a full data law of $\{Y(a, m), Y(a, 0), M, L, A\}$ which satisfies (1), (12) and (15) and marginalizes to the observed data law.

Remark. In view of assumption (1), Part (ii) of Theorem 3 is in principle restricted to functions $q_a(A, m, L)$ that satisfy $E\{q_a(A, m, L)|A, C\} = 0$ for all m. However, the user need not worry about choosing functions that satisfy this constraint as functions of only A, m and C vanish from expression (16).

Theorem 3 covers a much broader range of settings than is currently the case in the identification of population natural direct and indirect effects in which a time-dependent confounder L (a variable affected by treatment that in turn confounds the mediator-outcome relationship) renders natural direct and indirect effects unidentified (Avin et al., 2005). This is not only because assumption (12) relaxes (2), but also because no ignorability assumptions are required regarding the exposure-mediator association, as well as no assumptions about conditional independence of the counterfactuals $Y(a, m)$ and $M(a^*)$, besides assumption (12). Instead, Theorem 3 requires a priori knowledge of a selection-bias function $q_{a^*}(A, m, L)$ for the given a^*. In Section 4.3 we argue that the choice $q_{a^*}(A, m, L) = 0$ holds under a large class of plausible data-generating models that allow for exposure-induced confounding. Expression (16) with $q_{a^*}(A, m, L) = 0$ thus provides (parametric) identification results for natural direct and indirect effects in the exposed in the presence of such confounding. Our results moreover enable sensitivity analyses to assess the impact of departures from the assumption that the selection-bias function equals zero.

4.3 *The choice* $q_{a^*}(A, m, L) = 0$

By definition, the selection-bias function $q_{a^*}(A, m, L)$ can only differ from zero when there are mediator effects that vary with A. When these effects do not additionally vary with L (conditional on C), then $q_{a^*}(A, m, L)$ is a function of only A, m and C, which can be seen to vanish from expression (16). Thus for the investigator to choose a non-zero selection-bias function, there would essentially have to be a three way (additive) interaction between m, A and L, as we will demonstrate more formally in the next section. The selection-bias function can also be set to zero when there is no exposure-induced mediator-outcome confounding, for then either $Y(a^*, m) \perp\!\!\!\perp A|L$ so that $q_{a^*}(A, m, L) = 0$, or $M \perp\!\!\!\perp L|A, C$ in which case terms involving $q_{a^*}(A, m, L)$ vanish from expression (16). Finally, it can also be set to zero when L includes all causal risk factors of the outcome that are also associated with the mediator. Indeed, conditioning on all causal risk factors of the outcome (i.e., U in Figure 1) renders $Y(a, m)$ and A independent, conditional on L.

While it follows from Theorem 3, Part (ii), that the observed data can never provide evidence to support the choice $q_{a^*}(A, m, L) = 0$, we recommend using it as the primary reference choice when its plausibility cannot be rejected on subject-matter grounds (see the Web Appendix for alternative, natural choices). We make this recommendation for the following two reasons. First, it follows from the previous paragraph that $q_{a^*}(A, m, L) = 0$ under a class of data-generating mechanisms which is so broad, that choices other than zero might essentially be regarded as being in disagreement with the usual principles of parsimony. Indeed, in the hypothetical event that the data were informative about the selection-bias function, these principles would likely lead one to set $q_{a^*}(A, m, L)$ equal to zero since for $q_{a^*}(A, m, L)$ to be non-zero, there has to be a three way (additive) interaction between m, A and L, the evidence for which would typically be weak. Second, it follows from the first paragraph that the assumption, $q_{a^*}(A, m, L) = 0$, can be made more plausible by collecting data on causal risk factors of the outcome. We view this control which the investigator has over the plausibility of this assumption as a desirable characteristic and motive for adopting it as a 'reference assumption' (Vansteelandt, 2009a).

In the example, we could take $q_1(A, m, L) = 0$ when the effect of pre-eclampsia in the presence of adequate prenatal care is the same for women with the same smoking behavior, L, in the 2 prenatal care groups. For this choice, it follows from Theorem 3 that

$$\begin{aligned} E\left\{Y(1, M)|A = 0, C\right\} &= \int\int E(Y|M, L, A = 1)\left\{f(L|A = 1, C) - f(L|A = 0, C)\right. \\ &\quad \left. + f(L|M, A = 0, C)\right\} dL f(M|A = 0, C) dM. \qquad (17) \end{aligned}$$

4.4 *Sensitivity analysis*

Since by Part (ii) of Theorem 3, one cannot empirically verify whether $q_{a^*}(A,m,L)$ is a non-zero function of L, a zero selection-bias function should only be assumed if considered plausible. Whenever one has concerns about its plausibility, we recommend repeating the analysis for various choices of selection-bias functions.

To gain insight into the possible form of the selection-bias function, suppose that

$$E(Y|U,L,M,A) = h_0(L,M,A) + Uh_1(L,A) + Uh_2(L,M,A)$$

for arbitrary functions $h_0(L,M,A), h_1(L,A), h_2(L,M,A)$ satisfying $h_2(L,0,A)=0$, where U contains all prognostic factors of Y that are also associated with the intermediate confounder L (see Figure 1). Then it is easily shown under the assumptions of the causal diagram of Figure 1 that

$$q_a(A,m,L) = h_2(L,m,a)\left\{E(U|A,L) - E(U|A=a,L)\right\}.$$

Since, $E(U|A,L) \neq E(U|A=a,L)$ in the presence of intermediate confounding, the selection-bias function $q_a(A,m,L)$ differs from zero only when the mediator effect is modified (on the additive scale) by the unmeasured confounders U of the confounder-outcome association. Suppose in particular that $h_2(L,m,a)$ depends on L only through C in the sense that - with a slight abuse of notation - $h_2(L,m,a) = h_2(C,m,a)$. Suppose furthermore that (A,U,L) is multivariate normal so that $E(U|A,L) = \alpha_0 + \alpha_a A + \alpha_l L$. Then $q_a(A,m,L) = h_2(C,m,a)\alpha_a(A-a)$, which can be ignored since it does not vary conditional on A and C. It thus follows that under the above model, the selection-bias function will only differ from zero when either the mediator effect is modified by both U and L, or when the mediator effect is modified by U and, in addition, the association between U and A varies with L. Thus when $h_2(L,m,a) = \beta_0 m + \beta m L$ and $E(U|A,L) = \alpha_0 + \alpha A + \alpha_l L$, or $h_2(L,m,a) = \beta m$ and $E(U|A,L) = \alpha_0 + \alpha_a A + \alpha_l L + \alpha AL$, we have that

$$q_a(A,m,L) = \gamma m(A-a)L, \tag{18}$$

where $\gamma = \alpha\beta$. It is easily shown for this choice that $E\left\{Y(a, M)|A, C\right\}$ equals the value (17) obtained for $q_a(A, m, L) = 0$ plus

$$\gamma(A - a)\text{Cov}(M, L|A, C). \tag{19}$$

It follows that sensitivity of the results to the choice of selection-bias function will be weak when the degree of intermediate confounding is weak in the sense that M and L are weakly correlated conditional on exposure and baseline covariates. It further follows that from a computational point of view, a sensitivity analysis is straightforward as it merely requires modifying the estimate (17) by adding the contribution (19) for different choices of γ. The difficulty lies, however, in finding a plausible range of selection-bias functions or, in particular, of values of γ. In the next paragraph, we give some guidance as to how a realistic range might be chosen.

Let U be a scalar variate with mean zero and unit variance. Suppose furthermore that $h_2(L, m, a) = \beta_0 m + \beta m L$ and $E(U|A, L) = \alpha_0 + \alpha A + \alpha_l L$. Then $\beta = \text{SD}(Y)/\left\{\text{SD}(M)\text{SD}(L)\right\}$ would indicate a relatively large three way interaction between M, L and U and could therefore be taken as a maximum value in the sensitivity analysis. In the Web Appendix, we further show that if (A, U, L) follow a multivariate normal distribution, then

$$\alpha = \frac{-\tilde{R}_u \tilde{R}_a}{\text{SD}(A)(1 - \tilde{R}_a^2)},$$

where $\tilde{R}_u$ and $\tilde{R}_a$ are the root coefficients of determination corresponding to U and A, respectively, in the model for L. Here, $\tilde{R}_a$ is estimable from a regression of L on A because U and A are independent regressors under the causal diagram of Figure 1. Assuming for instance that the coefficient of determination for the model for the conditional mean of L, given A and U, does not exceed 0.9, one could choose the maximum value of α in the sensitivity analysis equal to $\pm(\tilde{R}_a\sqrt{0.9 - \tilde{R}_a^2})/(\text{SD}(A)(1 - \tilde{R}_a^2))$.

5. Data analysis

We analyze the 2003 US birth certificate data with adequate or inadequate prenatal care ($n = 2\,629\,247$; those with intermediate or superadequate care are excluded from the analysis for the purposes of this illustration) to evaluate direct and indirect effects of adequate or inadequate care (A) on the risk of preterm birth (Y) other than through pre-eclampsia (M). Adequacy of prenatal care categories are determined from data on the month prenatal care was initiated, on the number of visits and on gestational age according to the American College of Gynecologists recommendation as encoded in a modification of the APNCU inext (Kotelchuck, 1994; VanderWeele et al., 2009). The frequency of pre-eclampsia in the population is 3.16% and 2.30% in nonsmoking and smoking mothers with inadequate care, respectively, and 3.22% and 2.01% in nonsmoking and smoking mothers with adequate care, respectively. Furthermore, the frequency of smoking is 23.6% in mothers with inadequate care and 21.7% in mothers with adequate care. Our analysis considers mother's drinking, age category (below 20 years, between 20 and 35 years, or above 35 years), ethnicity (black, hispanic, native american, white), education and marital status as baseline confounders (C), and mother's smoking as an intermediate confounder (L), considering the possible beneficial effect of adequate care on smoking, which itself decreases the likelihood of pre-eclampsia. We fit a logistic regression model for the risk of preterm birth (Y) involving main effects of all variables and allowing for modification of the effect of pre-eclampsia (M) by adequate care and ethnicity (black), and for modification of the effect of smoking by drinking. We further fitted logistic regression models for the risk of pre-eclampsia involving main effects of all variables (except preterm birth) and allowing for modification of the effect of smoking by drinking, and for the risk of smoking involving main effects of all variables (except preterm birth and pre-eclampsia). Assuming that the selection-bias function is zero, the use of expression (17) resulted in $\hat{E}\{Y(1, M)|A = 0\} = 0.061$ after averaging out all baseline confounders. We thus estimate that the risk of preterm birth would have been 6.1% in those with inadequate care, had they received adequate care but their pre-eclampsia status remained unchanged. Since the observed risk of preterm birth in those with inadequate care is 12.4%, this corresponds with a natural direct effect for those with inadequate care of

$$\begin{aligned}\hat{E}\{Y(1,M)-Y|A=0\} &= \hat{E}\{Y(1,M(0))-Y(0,M(0))|A=0\} \\ &= -0.0639 \text{ (bootstrap 95\% CI -0.0648 tot -0.0630).}\end{aligned}$$

It thus follows that adequate care could decrease the risk of preterm birth in those with inadequate care with 6.4% other than by effecting pre-eclampsia. Had these women received adequate care, the additional indirect effect of adequate care through pre-eclampsia would be in the opposite direction, but negligible:

$$\begin{aligned}\hat{E}\{Y(1)-Y(1,M)|A=0\} &= \hat{E}\{Y(1,M(1))-Y(1,M(0))|A=0\} \\ &= 0.00020 \text{ (bootstrap 95\% CI 0.00015 to 0.00025).}\end{aligned}$$

That this effect is small is perhaps not surprising since the effect of adequate care on pre-eclampsia by decreasing smoking would likely increase pre-eclampsia whereas the effect of adequate care not through smoking would likely be in the other direction, decreasing pre-eclampsia.

We subsequently allowed for a sensitivity analysis by varying the choice of selection bias function $q_1(1,m,L)$. Note that $q_1(1,0,L)=0$ by definition and thus that only specification of $q_1(1,1,L)$ is required (where $m=1$ refers to pre-eclampsia). By definition, $q_1(1,1,L)$ contrasts the effect of pre-eclampsia in the presence of adequate care between those without and with adequate care and given smoking behavior S and baseline covariates C (where $L=(C,S)$). In a sensitivity analysis, we assumed that $q_1(1,1,L)=q(C)+\lambda S$, where the choice of $q(C)$ is irrelevant as it cancels from expression (16), and where λ was varied from -0.1 to 0.1 which is a wide range since the risk of preterm birth even in the inadequate care group is only 0.124. Interestingly, the result in Figure 2 reveals virtually no sensitivity of the direct and indirect effect estimates to the amount of selection bias.

Our analysis here is intended only as an illustration, as it is a simplification of what is a more complex reality and therefore limited in the substantive conclusions that can be drawn. Both pre-eclampsia and preterm birth are time-varying process and here we have treated them as dichotomous, which may induce bias (Zhang et al., 2011). Furthermore, data is only available for pre-eclampsia as a dichotomous variable; moreover not all pre-eclampsia is diagnosed. Current research (Ogburn and VanderWeele, 2012) indicates that under many conditions, dichotomization of a mediator will lead to overestimates of the direct effect and underestimates of the indirect effect (though this intuition can sometimes fail). In this example, dichotomization of the mediator pre-eclampsia (which was the only measure available in this dataset) may have led to an underestimation of the mediated effect.

Another limitation of our analysis is that the mediator M is only partially manipulable and different potential interventions to address pre-eclampsia may well have different effects on the outcome, preterm birth. This issue is sometimes referred to as a problem of "multiple versions" and makes the interpretation of effect estimates somewhat more difficult; in some circumstances such effects can be interpreted as the consequence of an intervention that randomly selects, conditional on the covariates, a "version of treatment", from the distribution of that which actually occurred in the population (Hernan and VanderWeele, 2011).

[Figure 2 about here.]

6. Discussion

Controlled direct effects can be identified in the presence of exposure-induced mediator-outcome confounding, but have a more limited utility because they cannot be used for decomposition of a total effect into direct and indirect effects, and because of their more stringent interpretation which involves fixing the mediator at the same value uniformly in the population. The formalism of natural direct and indirect effects remedies these limitations, but requires stronger identification conditions which essentially preclude such intermediate confounding. We have focused on the identification of natural direct and indirect effects on the exposed, which combine the best of both formalisms: they enable decomposition of a total effect, allow for natural variation in the level at which the mediator is controlled, and are essentially identifiable under the same conditions as for controlled direct effects. Indeed, although our formalism demands additional knowledge on a selection-bias function, we have shown that it evaluates to zero under a large class of data-generating mechanisms.

We have shown that natural direct and indirect effects on the exposed coincide with the corresponding population-averaged natural effects under specific ignorability assumptions which are in particular satisfied when the exposure is randomly assigned. Thus, under these conditions, our development moreover provides a method for identifying population natural direct and indirect effects in the presence of exposure-induced mediator-outcome confounding, which existing methodology has not been able to address (Avin et al., 2005). Our effect definitions are thus theoretically appealing. They are moreover the effects that are of substantive interest whenever the effect of treatment on the treated (rather than the population averaged treatment effect) is in view and one wants to further examine the mechanisms responsible for this effect.

That natural direct and indirect effects on the exposed are identifiable under weaker conditions than the corresponding population averaged effects should not come as a surprise. Similar results are for instance found in the literature on instrumental variables (Hernan and Robins, 2006). The reason is that by focussing solely on the exposed, one is less ambitious about the causal inferences one attempts to make as one generally loses the ability to make claims about the effects of arbitrary interventions on A. Note therefore that the natural direct effect on the exposed $\psi = E\{Y - Y(a^*, M)|A = a\} = E\{Y(a, M(a)) - Y(a^*, M(a))|A = a\}$ must be cautiously interpreted as it only states that setting A to a^* within those with $A = a$, while holding the mediator fixed, will change the expected outcome by ψ. However, it does not enable one to infer the effect of an $a - a^*$ unit increase in the exposure in the general population as subgroups with different observed exposures might experience different effects. In further analogy to the instrumental variables setting, the identification results that we provide (corresponding to zero selection-bias function) are parametric in the sense that they are restricted to data-generating models that exclude additive three-way interactions between exposure, mediator and exposure-induced confounders. When such interactions are anticipated, our results still enable a study of the sensitivity to deviations away from that assumption. Like the sensitivity analysis results of Tchetgen Tchetgen and Shpitser (2011), we thus allow for exposure-induced confounding, but in contrast, provide adjustment for measured exposure-induced confounders so that less sensitivity is expected with our approach.

In the application, we made use of substitution estimators, obtained by substituting the observed data distribution by a consistent estimator. In more general settings, such substitution estimators may not have a desirable performance for various reasons. First, because functionals like (16) involve high-dimensional integration, substitution estimators can be computationally tedious to obtain. Second, the modeling of the observed data distribution, which may involve high-dimensional confounders L, can make these estimators greedy in demanding parametric modeling assumptions. Further subtleties arise because parsimonious models for the observed data distribution need not translate into parsimonious models for the direct and indirect effect. This may not only make the results unattractive for reporting, but also make interesting hypotheses difficult to test. In future work, recourse to alternative strategies will be sought based on direct modeling (van der Laan and Petersen, 2008) and estimation of natural direct and indirect effects in the exposed. We expect this will be quite feasible in the absence of intermediate confounding where Theorem 3 suggests that progress can be made using inverse probability weighting. This may be more challenging in the presence of intermediate confounding because of the possibly high-dimensional confounder distributions appearing in (16).

Finally, Avin et al. (2005) showed that a time-dependent confounder L renders population natural direct and indirect effects unidentified. Our results have shown that progress can still be made when the exposure is randomly assigned, and more broadly for exposure effects on the exposed. It remains to be studied how the results of Avin et al. (2005) on path-specific effects in causal diagrams that involve multiple mediators, extend to our formalism.

Supplementary Materials

The Web Appendix referenced in Section 4.3 is available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

The first author acknowledges support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy). The second author acknowledges funding from the National Institutes of Health grant HD060696.

References

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

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51,** 1173–1182.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *Int. J. Epidem.* **31,** 163–165.

Didelez, V., Dawid, A., and Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence*, pages 138–146.

Hernan, M. and Robins, J. (2006). Instruments for causal inference - An epidemiologist's dream? *Epidemiology* **17,** 360–372.

Hernan, M. A. and VanderWeele, T. J. (2011). Compound Treatments and Transportability of Causal Inference. *Epidemiology* **22,** 368–377.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science* **25,** 51–71.

Joffe, M., Byrne, C., and Colditz, G. (2001). Postmenopausal hormone use, screening, and breast cancer: Characterization and control of a bias. *Epidemiology* **12,** 429–438.

Kotelchuck, M. (1994). An evaluation of the Kessner adequacy of prenatal care index and a proposed adequacy of prenatal care utilization index. *American Journal of Public Health* **84,** 1414–1420.

Kurth, T., Walker, A., Glynn, R., Chan, K., Gaziano, J., Berger, K., and Robins, J. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* **163,** 262–270.

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

Ogburn, E. and VanderWeele, T. (2012). Analytic results on the bias due to nondifferential misclassification of a binary mediator. *Epidemiology, in press* .

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on*

Uncertainty and Artificial Intelligence, pages 411–420, San Francisco. Morgan Kaufmann.

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

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In *Computation, causation, and discovery,* pages 349–405. AAAI Press, Menlo Park, CA.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In Green, P., Hjort, N., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 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.

Rosenblum, M., Jewell, N. P., van der Laan, M., Shiboski, S., van der Straten, A., and Padian, N. (2009). Analysing direct effects in randomized trials with secondary interventions: an application to human immunodeficiency virus prevention trials. *Journal of the Royal Statistical Society, Series A* **172,** 443–465.

Tchetgen Tchetgen, E. and Shpitser, I. (2011). Semiparametric estimation of models for natural direct and indirect effects. Technical report, Harvard University.

van der Laan, M. and Petersen, M. (2008). Direct effect models. *Int. J. Biostat.* **4,** Article 23.

VanderWeele, T. (2009). Mediation and mechanism. *European Journal of Epidemiology* **24,** 217–224.

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

VanderWeele, T. J., Lantos, J. D., Siddique, J., and Lauderdale, D. S. (2009). A comparison of four prenatal care indices in birth outcome models: Comparable results for predicting small-for-gestational-age outcome but different results for preterm birth or infant mortality. *Journal of Clinical Epidemiology* **62,** 438–445.

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

Vansteelandt, S. (2009a). Discussion on "Identifiability and Estimation of Causal Effects in Randomized Trials with Noncompliance and Completely Nonignorable Missing Data". *Biometrics* **65,** 686–689.

Vansteelandt, S. (2009b). Estimating direct effects in cohort and case-control studies. *Epidemiology* **20,** 851–860.

Vansteelandt, S., Mertens, K., Suetens, C., and Goetghebeur, E. (2009). Marginal structural models for partial exposure regimes. *Biostatistics* **10,** 46–59.

Zhang, M., Joffe, M. M., and Small, D. S. (2011). Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics* **39,** 131–173.

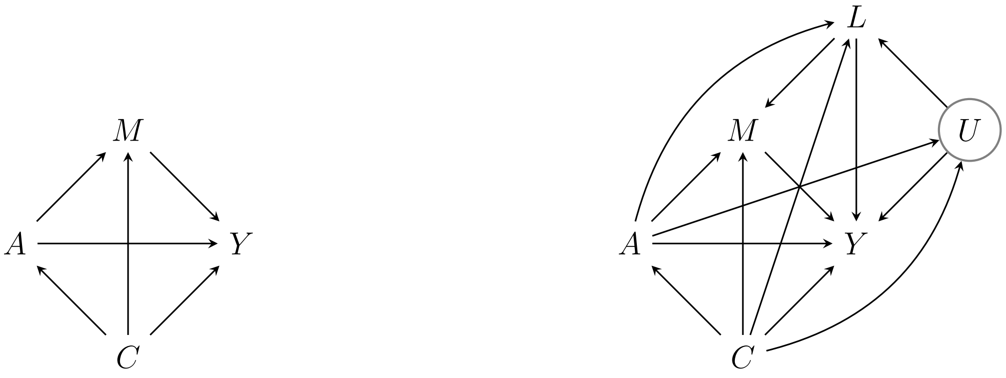

Figure 1. Left: Causal diagram without exposure-induced confounding. Right: Causal diagram with exposure-induced confounding.

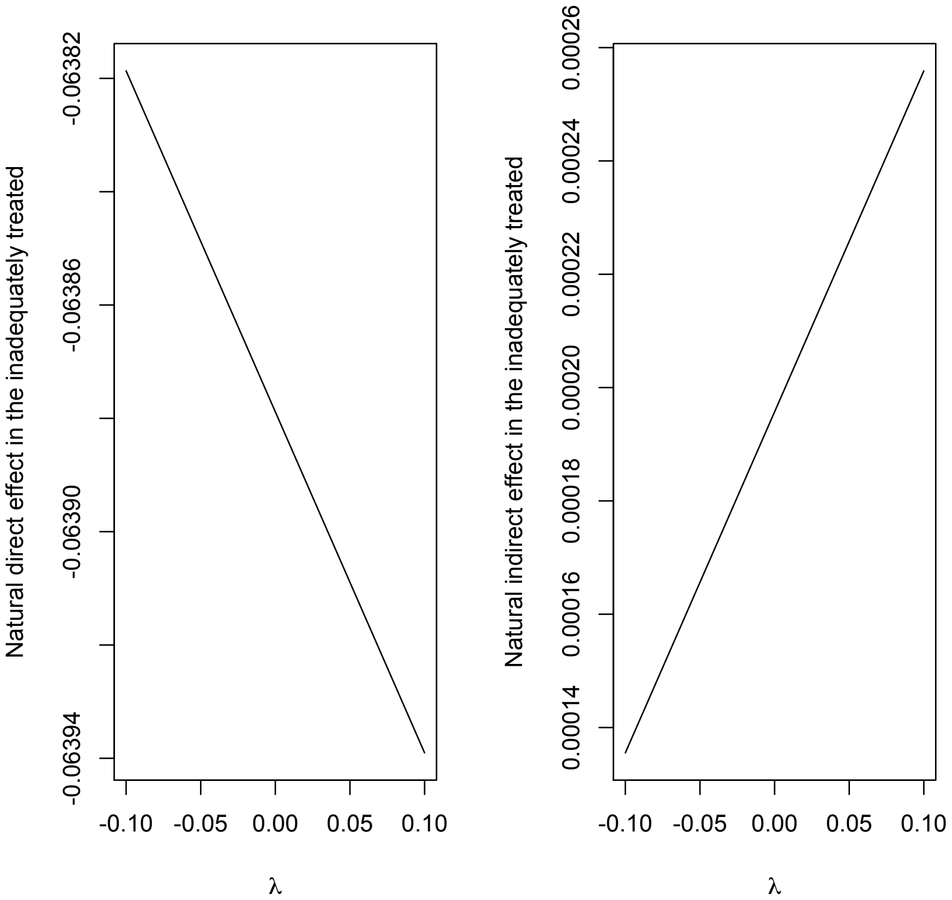

Figure 2. Natural direct (left) and indirect (right) effect of adequate care on the risk of preterm birth other than through pre-eclampsia in the inadequately treated.