Statistics and Its Interface Volume 2 (2009) 457–468

Conceptual issues concerning mediation, interventions and composition

Tyler J. VanderWeele* and Stijn Vansteelandt

Concepts concerning mediation in the causal inference literature are reviewed. Notions of direct and indirect effects from a counterfactual approach to mediation are compared with those arising from the standard regression approach to mediation of Baron and Kenny (1986), commonly utilized in the social science literature. It is shown that concepts of direct and indirect effect from causal inference generalize those described by Baron and Kenny and that under appropriate identification assumptions these more general direct and indirect effects from causal inference can be estimated using regression even when there are interactions between the primary exposure of interest and the mediator. A number of conceptual issues are discussed concerning the interpretation of identification conditions for mediation, the notion of counterfactuals based on hypothetical interventions and the so called consistency and composition assumptions.

Keywords and phrases: Causal inference, Counterfactuals, Direct and indirect effects, Intervention, Mediation.

The notion of mediation concerns the extent to which the effect of one variable on another is mediated by some possible intermediate variable. As such, notions of mediation concern causality. The use of mediation analysis has become quite common in the social sciences. An approach based on regression analysis advocated by Baron and Kenny (1986) is now utilized routinely, especially within the literature on psychology. More recently, an approach to mediation arising from the causal inference literature and based on the notion of counterfactuals has been proposed (Robins and Greenland, 1992; Pearl, 2001). The current paper is structured in two parts. In the first part we will review notions of direct and indirect effects from the causal inference literature on mediation and will relate these notions to the approach advocated by Baron and Kenny (1986). We will show how the notions from causal inference generalize those arising from the Baron and Kenny approach. In particular, under certain identification conditions, the approach based on causal inference allows for the definition of direct and indirect effects and for effect decomposition of a total effect into a direct and indirect effect even in settings with interaction and nonlinearities. Under appropriate identification conditions, the direct and indirect effects defined in the causal inference literature can also be estimated in a regression framework. In the second part of the paper we will consider in some detail a number of conceptual issues raised by the causal inference or counterfactual approach to mediation. We will discuss the constraints imposed by identification conditions for mediation; we will discuss the extent to which an approach to mediation based on interventions is possible and we will finally consider the interpretation of a conceptual assumption, sometimes referred to as "composition", which is made in the causal inference work on mediation when decomposing total effects into direct and indirect effects.

*Corresponding author.

DIRECT AND INDIRECT EFFECTS IN CAUSAL INFERENCE

To formalize the meaning of a direct effect, we conceptualize for each subject the existence of a so-called counterfactual outcome $Y(a)$, which denotes the outcome that we would — possibly contrary to fact — have observed for that subject had the exposure A been set to the value a through some intervention or manipulation (Rubin, 1978; Hernán, 2004). Variables such as $Y(a)$ are referred to as "potential outcomes" or "counterfactual outcomes". If the exposure A is dichotomous (e.g., taking values 0 for no exposure and 1 otherwise), then we can think of each subject having 2 such counterfactual outcomes, $Y(0)$ and $Y(1)$. The *(average) causal effect* of the exposure on the outcome can then be defined as the expected difference $E[Y(1)-Y(0)]$ between both counterfactual outcomes for the same study population. This is to be contrasted with the more usual expected difference $E[Y|A=1]-E[Y|A=0]$, where Y denotes the observed outcome, which may not carry the interpretation of a causal effect when the subgroups of exposed and unexposed subjects are not inherently comparable. More generally, we define the *conditional causal effect* of exposure level a versus 0 (where we let 0 denote an arbitrary reference level) on the outcome, given pre-exposure covariates C, as the expected contrast $E[Y(a)-Y(0)|C]$.

To be able to identify total causal effects, the following 2 assumptions must be made. The so-called consistency assumption states that amongst subjects with observed exposure level $A=a$, the observed outcome Y is equal to the potential outcome $Y(a)$ (i.e., $Y(a)=Y$ when $A=a$). Under this assumption, we can observe one of the potential

outcomes for each subject, namely the one corresponding to the observed exposure level (i.e., $Y = Y(A)$). We return to this assumption below. We need one further assumption for the identification of total causal effects; we also need some additional notation. For random variables A, B and C, let $A \perp\!\!\!\perp B|C$ denote that A is conditionally independent of B, given C. For the identification of total causal effects, we will assume that subjects with different observed exposure levels A, but the same pre-exposure characteristics C, are comparable in the sense that

$$Y(a) \perp\!\!\!\perp A|C$$

for all exposure levels a. This assumption states that a subject's choice of exposure level A, while possibly related to pre-exposure characteristics C, has no residual dependence on how that subject would fare under an arbitrary, fixed exposure level. It is usually referred to as the *no unmeasured confounders assumption* as it effectively states that the variables in C are the only confounders of the association between exposure and outcome. Both these assumptions cannot be tested on the basis of the observed data, but, in combination, are sufficient for identifying the conditional causal effect as

$$\begin{aligned} &E[Y(a) - Y(0)|C] \\ &\quad = E[Y(a)|C] - E[Y(0)|C] \\ &\quad = E[Y(a)|A = a, C] - E[Y(0)|A = 0, C] \\ &\quad = E[Y|A = a, C] - E[Y|A = 0, C]. \end{aligned}$$

We could take averages over the distribution of C to obtain average causal effects, $E[Y(1) - Y(0)]$.

By extending the previous concepts to a joint exposure (A, M) where M is the potential mediator, definitions of direct and indirect effects can be constructed. For each subject, let us define $Y(a, m)$ to be the outcome that we would — possibly contrary to fact — have observed for that subject had the exposure A been set to the value a and, likewise, M to the value m, through some intervention or manipulation. Similarly, we can consider counterfactual variables $M(a)$ which denote the value of the mediator if — possibly contrary to fact — the exposure A were set to a. For a dichotomous exposure, the *controlled direct effect* of the exposure on the outcome, controlling for M, can then be defined as the expected contrast $E[Y(1, m) - Y(0, m)]$ (Robins and Greenland, 1992; Pearl, 2001). It expresses the exposure effect that would be realized if the mediator were controlled at level m uniformly in the population. For instance, in accordance with Figure 1, let A be the father's occupation, let Y be the respondent's income, let M be the respondent's occupation, let C_1 be the father's education, and let C_2 be the respondent's education. Then $E[Y(a, m) - Y(0, m)]$ expresses the average change in income that would be realized in a subgroup of respondents if their father changed occupation (from 0 to a), but their own occupation were kept uniformly at level m. More generally, we will define the *conditional controlled direct effect* of exposure level a versus 0 on the outcome (controlling for M), given covariates C, as the expected contrast $E[Y(a, m) - Y(0, m)|C]$.

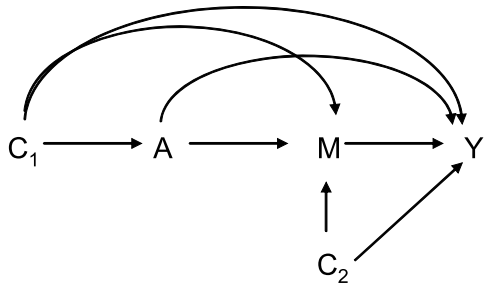

Figure 1. Example of the effect of A on Y mediated by M with both exposure-outcome confounders and mediator-outcome confounders.

The consistency assumption for joint exposure (A, M) is then that amongst the subgroup with observed exposure $A = a$ and observed mediator $M = m$, the observed outcome Y is equal to $Y(a, m)$. The consistency assumption for the effect of the exposure on the mediator is that amongst the subgroup with observed exposure $A = a$ the observed mediator M is equal to $M(a)$. The assumption of no-unmeasured-confounders for the exposure-outcome relationship can then be expressed as

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

for all levels of a and m. However, controlled direct effects in general require stronger conditions for identification than do total causal effects. This is because the definition of a controlled direct effect requires evaluating the impact of holding the mediator M fixed. For this purpose, one must know all confounders of the association between mediator and outcome, which we formally express as

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

for all levels of a and m. To identify controlled direct effects, the set C must contain all of the confounders of both the exposure-outcome relationship and the mediator-outcome relationship i.e. in Figure 1 control must be made for both C_1 and C_2 to identify controlled direct effects. If control is not made for the variables that confound the relationship between the mediator and the outcome (the variables C_2 in Figure 1) then estimates of direct effects will generally be biased (Cole and Hernán, 2002). In the early mediation literature, this point about controlling for the mediator-outcome confounders was made by Judd and Kenny (1981) but was not pointed out by Baron and Kenny (1986) and was also subsequently ignored by much of the social science literature on mediation. The importance of controlling for the confounders of the mediator-outcome relationship has been

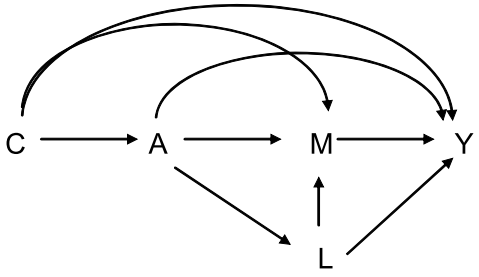

Figure 2. Example of an effect of the exposure confounding the mediator-outcome relationship.

emphasized in the causal inference literature on mediation (Robins and Greenland, 1992; Pearl, 2001; Cole and Hernán, 2002). If assumptions (1) and (2) hold, then controlled direct effects are identified (Pearl, 2001) by

$$E[Y(a,m) - Y(0,m)|C] \\ = E[Y|A=a, M=m, C] - E[Y|A=0, M=m, C].$$

We could take averages over the distribution of C to obtain average controlled direct effects, $E[Y(a,m) - Y(0,m)]$. Note that, because the mediator M arises later in time than the exposure A, it is possible that the exposure itself may affect one or more of the confounders of the mediator-outcome relationship. Such an example is given in Figure 2 with L being both an effect of exposure A and a confounder of the mediator-outcome relationship. Note that in Figure 2, L could also be considered to be a mediator of the effect of A on Y; which variable is taken to be the mediator of interest will depend on the context. In this paper we will use M to denote the mediator of interest and L to denote any confounders of the mediator-outcome relationship that are affected by the exposure A. When there are confounders of the mediator-outcome relationship that are affected by the exposure, condition (2) can be modified to

$$Y(a,m) \perp\!\!\!\perp M|A, C, L$$

for all levels of a and m and then controlled direct effects can still be identified (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2009a). For most of this paper, however, we will restrict our attention to settings in which the confounders of the mediator-outcome relationship are not affected by exposure. We will consider the scope of such settings further below.

There are a number of limitations to the concept of a controlled direct effect. First, as in the previous example, it is often not realistic to imagine scenarios where one would consider forcing the mediator to be the same for all subjects in the population (e.g. forcing all subjects to have the same occupation). Second, indirect effects cannot be defined in a similar manner as controlled direct effects because it is impossible to hold a set of variables fixed, in such a way that the effect of exposure on outcome would circumvent the direct pathway. In particular, the total causal effect, say $E[Y(a) - Y(0)]$, minus the controlled direct effect, say $E[Y(a,m) - Y(0,m)]$, may not carry the interpretation of an indirect effect. Because of potential exposure-mediator interactions, the difference may be non-zero even if the exposure has no effect on the mediator so that none of the effect of the exposure is mediated by M (Kaufman et al., 2004; VanderWeele 2009b). Both limitations can be overcome by considering so-called *natural direct effects* (Pearl, 2001; Robins, 2003) which may be defined as the expected contrast $E[Y(a, M(0)) - Y(0, M(0))]$. Robins and Greenland (1992) refer to this quantity as the *pure direct effect* to distinguish it from the *total direct effect*, $E[Y(a, M(a)) - Y(0, M(a))]$; both are instances of what Pearl (2001) calls natural direct effects and we will thus use the terms *pure natural direct effect* and *total natural direct effect*. The pure natural direct effect, $E[Y(a, M(0)) - Y(0, M(0))]$, expresses the effect that would be realized if the exposure were administered, but its effect on the mediator were somehow blocked, or equivalently, if the mediator were kept at the level it would have taken in the absence of the exposure. More generally, we will define the *conditional pure natural direct effect* of exposure level a versus 0 on the outcome (other than through modifying M), given pre-exposure covariates C, as the expected contrast $E[Y(a, M(0)) - Y(0, M(0))|C]$. In the context of the example, this expresses the average change in income that would be realized in a subgroup of respondents (all of whose fathers had the same education) if their father changed occupation (from 0 to a), but they kept their own occupation. While controlled direct effects are often of greater interest in policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms (Robins, 2003; Joffe et al., 2007).

In considering natural direct effects, one typically makes not only the consistency assumption but also a composition assumption that $Y(a) = Y(a, M(a))$ i.e. that the potential outcome $Y(a)$ intervening to set A to a is equal to the potential outcome $Y(a, M(a))$ intervening to set A to a and to set M to the value it would have been if A had been a. Under this composition assumption the pure natural direct effect $E[Y(1, M(0)) - Y(0, M(0))|C]$ can also be expressed as $E[Y(1, M(0)) - Y(0)|C]$. We return to the interpretation of the composition assumption below.

The use of natural direct effects overcomes the previously listed limitations of controlled direct effects. First, this is so because the level $M(0)$ at which the mediator is controlled allows for natural variation between subjects. Second, this is so because the difference between the total causal effect and a pure natural direct effect

$$E[Y(a) - Y(0)|C] - E[Y(a, M(0)) - Y(0)|C] \\ = E[Y(a, M(a)) - Y(a, M(0))|C]$$

expresses how much the outcome would change on average if the exposure were controlled at level a, but the mediator were changed from level $M(0)$ to $M(a)$. It thus carries the interpretation of an indirect effect and will be termed the *total natural indirect effect* (Robins and Greenland, 1992; Robins, 2003). Importantly, the above effect decomposition does not assume that the functional form relating A, M, and Y is linear nor that there is no interaction between the effects of A and M on Y. Likewise, the difference between the total effect and the total natural direct effect, $E[Y(a, M(a)) - Y(0, M(a))|C] = E[Y(a) - Y(0, M(a))|C]$, gives

$$E[Y(a) - Y(0)|C] - E[Y(a) - Y(0, M(a))|C] = E[Y(0, M(a)) - Y(0, M(0))|C],$$

which expresses how much the outcome would change on average if the exposure were controlled at level 0, but the mediator were changed from its natural level $M(0)$ to the level $M(a)$ which it would have taken at exposure level a. This is termed the *pure natural indirect effect* (Robins and Greenland, 1992; Robins, 2003). In the context of the example, this expresses the average change in income that would be realized in a subgroup of respondents (all of whose fathers had the same education) if their father's occupation were uniformly controlled at the reference level 0, but they changed their occupation to what they would have had if their father had a different occupation a.

Controlled and natural direct and indirect effects can all be defined so as to compare levels of exposure a and a^* rather than a and 0. Thus, the controlled direct effect under this comparison is $E[Y(a, m) - Y(a^*, m)|C]$; the pure natural direct effect is $E[Y(a, M(a^*)) - Y(a^*, M(a^*))|C]$; the total natural indirect effect is $E[Y(a, M(a)) - Y(a, M(a^*))|C]$; the total natural direct effect is $E[Y(a, M(a)) - Y(a^*, M(a))|C]$; and the pure natural indirect effect is $E[Y(a^*, M(a)) - Y(a^*, M(a^*))|C]$.

Identification of natural direct effects, like controlled direct effects, requires assumptions (1) and (2) but relies on additional assumptions as well. First, to be able to assess what values the mediator would take if the exposure were controlled, one must have measured all confounders of the association between exposure and mediator, which we formally express as

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

for all levels of a. In addition to assumptions (1)–(3) the identification of natural direct and indirect effects is generally made by imposing an additional assumption. Pearl (2001), for example, effectively identifies natural direct and indirect effects by also assuming

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

for all levels of a, a^* and m. For example, for the natural direct effect $Y(a, M(0)) - Y(0, M(0))$ we would need (4) to hold for $a^* = 0$. Condition (4) is somewhat difficult to interpret but will generally hold if condition (2) holds and if there are no variables L that are effects of exposure and that confound the mediator-outcome relationship (Pearl, 2001). Under assumptions (1)–(4), the natural direct effect can be identified (Pearl, 2001) by

$$\begin{aligned} E[Y(a, M(0)) - Y(0, M(0))|C] &= \int \{E[Y|A = a, M = m, C] \\ &\quad - E[Y|A = 0, M = m, C]\} \\ &\quad \times f(M = m|A = 0, C) dm. \end{aligned} \tag{5}$$

Natural indirect effects can be obtained by subtracting a natural direct effect from a total effect as explained above. Note that even in settings in which the identification conditions for controlled and natural direct and indirect effects are not satisfied, data can sometimes still be used to obtain bounds on these effects (Kaufman et al., 2005, 2009; Cai et al., 2008; Sjölander, 2009).

Alternatively, instead of assumption (4), one can identify natural direct and indirect effects by making the no-interaction assumption that $Y(a, m) - Y(0, m)$ is a random variable that does not depend on m. This assumption states that the effect of changing the exposure from 0 to a while holding the mediator fixed, is the same no matter what value m to which one intervenes to fix the mediator. Under assumptions (1) and (2) and the no-interaction assumption, natural direct effects can be identified (Robins, 2003). Petersen et al. (2006) consider an assumption that essentially allows for identification of natural direct and indirect effects under assumptions (1)–(3) along with a disjunction of assumption (4) and the no-interaction assumption. In the next section, we will show how the approach of Baron and Kenny (1986) of using regression to assess mediation can be extended to settings including interactions. Instead of the no-interaction assumption, we will thus assume that conditions (1)–(4) hold. That is, we will restrict our consideration to settings in which there are no variables L that are effects of exposure and that confound the mediator-outcome relationship and we will assume the set C contains all confounders of the exposure-outcome, mediator-outcome and exposure-mediator relationships.

DIRECT AND INDIRECT EFFECTS USING REGRESSION

In assessing mediation, Baron and Kenny (1986) proposed using regression models such as

$$E[M|A = a, C = c] = \beta_0 + \beta_1 a + \beta_2' c \tag{6}$$

and

$$E[Y|A = a, M = m, C = c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c \tag{7}$$

in order to assess mediation. They proposed that the direct effect be assessed by estimating θ_1 and that the indirect effect be assessed by estimating $\beta_1\theta_2$; subsequent literature has discussed a variety of estimation procedures (MacKinnon, 2008). Importantly, the regression models Baron and Kenny proposed do not include an interaction term $\theta_3 am$ in the regression model for Y. In this section we show how the notions of direct and indirect effects from causal inference introduced in the previous section can extend the Baron and Kenny approach in settings in which there is an interaction term $\theta_3 am$ in the regression model for Y i.e. in settings in which the exposure and the mediator interact in their effects on the outcome. We assume that the mediator M and outcome Y are continuous; the results will apply to arbitrary exposures A. When the mediator M is dichotomous rather than continuous a somewhat similar approach to the one described here could potentially be used; however, if the model for M is logistic, rather than linear, then the analytic formulas for the natural direct and indirect effects no longer take quite as simple a form.

Suppose then that, instead of using regression models (6) and (7), one were to use regression model (6) and the following regression model

$$E[Y|A=a,M=m,C=c] = \theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c \tag{8}$$

i.e. one were to use the regression models of Baron and Kenny (1986) but to include a product term $\theta_3 am$ in the regression model for Y. In the appendix we show that under assumptions (1)–(4) and under correct specification of the regression models, the controlled direct effect and the natural direct and indirect effects are given by:

$$E[Y(a,m)-Y(a^*,m)|C] = (\theta_1+\theta_3 m)(a-a^*) \tag{9}$$

$$E[Y(a,M(a^*))-Y(a^*,M(a^*))|C] = (\theta_1+\theta_3\beta_0+\theta_3\beta_1 a^*+\theta_3\beta_2' C)(a-a^*)$$

$$E[Y(a,M(a))-Y(a,M(a^*))|C] = (\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*).$$

Note that the expression for the controlled direct effect only requires assumptions (1) and (2) and that the regression model (8) is correctly specified; the expressions for natural direct and indirect effects requires assumptions (1)–(4) and correct specification of models (6) and (8).

Several points merit attention. First, if there were indeed no product term $\theta_3 am$ in the regression model, i.e. if $\theta_3=0$, then the expressions in (9) reduce to

$$E[Y(a,m)-Y(a^*,m)|C] = \theta_1(a-a^*)$$
$$E[Y(a,M(a^*))-Y(a^*,M(a^*))|C] = \theta_1(a-a^*)$$
$$E[Y(a,M(a))-Y(a,M(a^*))|C] = \theta_2\beta_1(a-a^*)$$

and thus the controlled and the natural direct effects coincide and are equal to θ_1 for a one unit change in a, which is the direct effect given by the Baron and Kenny approach; the natural indirect effect is equal to $\theta_2\beta_1$ for a one unit change in a, which thus coincides with the indirect effect given by the Baron and Kenny approach. Thus when $\theta_3=0$, the notions of direct and indirect effects from causal inference reduce to the effects given in Baron and Kenny's proposal.

Second, when there is interaction so that $\theta_3\neq 0$, one can still define direct and indirect effects as described above and one can still decompose a total effect into a natural direct effect and a natural indirect effect; one can moreover, still use regression models to estimate natural direct and indirect effects and doing so gives the expressions given in (9) provided assumptions (1)–(4) hold. The definitions of natural direct and indirect effects given in the causal inference literature thus usefully extend concepts of direct and indirect effects from the social science literature to include settings with interactions. The definitions from causal inference given above furthermore apply also to settings with non-linear models (cf. Pearl, 2001; van der Laan and Petersen, 2008; VanderWeele, 2009a).

Third, the expressions given in (9) are for the pure natural direct effect, $E[Y(a,M(a^*))-Y(a^*,M(a^*))|C]$, and the total natural indirect effect, $E[Y(a,M(a))-Y(a,M(a^*))|C]$. One can similarly obtain expressions for the total natural direct effect, $E[Y(a,M(a))-Y(a^*,M(a))|C]$, and the pure natural indirect effect, $E[Y(a^*,M(a))-Y(a^*,M(a^*))|C]$. Using calculations similar to those in the appendix one obtains

$$E[Y(a,M(a))-Y(a^*,M(a))|C] = (\theta_1+\theta_3\beta_0+\theta_3\beta_1 a+\theta_3\beta_2' C)(a-a^*)$$
$$E[Y(a^*,M(a))-Y(a^*,M(a^*))|C] = (\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*).$$

The expression for the total natural direct effect differs from pure natural direct effect given in (9) in that it has the term $\theta_3\beta_1 a$, rather than $\theta_3\beta_1 a^*$. The expression for the pure natural indirect effect differs from total natural indirect effect given in (9) in that it has the term $\theta_3\beta_1 a^*$ rather than $\theta_3\beta_1 a$.

Finally, if the formulas in (9) are used to estimate direct and indirect effects by using estimates from linear regressions (6) and (8) then standard errors for the estimates of these direct and indirect effects could be obtained either through bootstrapping (Efron and Tibshirani, 1993; MacKinnon, 2008) or by the delta method as described in the appendix.

In this section we have restricted attention to the setting in which there are no effects of exposure that confound the mediator-outcome relationship. Some progress can be made even in settings in which there are such exposure-induced mediator-outcome confounders but estimation in such cases requires techniques beyond simple regression analysis (Robins, 1999; van der Laan and Petersen, 2008;

Goetgeluk et al., 2008; VanderWeele, 2009a; Vansteelandt, 2009). In the next section we will consider in greater detail this assumption that there are no effects of exposure that confound the mediator-outcome relationship.

CONCEPTUAL ISSUES CONCERNING THE IDENTIFICATION OF NATURAL DIRECT AND INDIRECT EFFECTS

In the previous section, to derive the expressions for direct and indirect effects given in (9) we have assumed that condition (4) would hold. Condition (4) allowed for the identification of natural direct and indirect effects but condition (4) is a strong assumption. Pearl (2001) gives a graphical interpretation of condition (4) essentially showing that it requires that there be no consequence of exposure that confounds the relationship between the mediator and the outcome. Condition (4) will be violated if there are such exposure-induced confounders, irrespective of whether or not data is available on them. Condition (4) contrasts with conditions (1)–(3). Conditions (1)–(3) could potentially be satisfied, at least approximately, by collecting data on more and more confounding variables C; conditions (1)–(3) are no-*unmeasured*-confounding assumptions. Condition (4) however requires that there be no effect of exposure that confounds the relationship between the mediator and the outcome; if there are such variables condition (4) will be violated irrespective of whether data is available for all such variables.

In many contexts, condition (4) will not be reasonable. Thus in the example concerning the effects of a father's occupational choice, A, on a respondent's income, Y, as mediated by respondent's occupation, M, if the father's occupational choice affects the respondent's education level, which we might denote by L in Figure 2, then condition (4) will be violated. If the father's occupational choice does not affect the respondent's education level, as in Figure 1 where we would denote respondent's education level by C_2, then condition (4) may be satisfied. In general one can imagine many variables on the pathway between the exposure of interest and the mediator. Condition (4), that there are no effects of exposure that confound the mediator-outcome relationship, essentially then requires that of all the variables that are on the pathway from the exposure to the mediator, none of these also affects the outcome. As noted above, this assumption will often be violated and in such cases, natural direct and indirect effects will not in general be identified. There are exceptions; for example, natural direct and indirect effects can be identified as in (5) without assumption (4) if the no-interaction assumption holds (Robins, 2003); see Petersen et al. (2006), Hafeman and VanderWeele (2009) and Imai et al. (2009) for other exceptions. However, it has been shown that whenever there is a consequence of the exposure that also affects both the mediator and the outcome then natural direct and indirect effects are not in general identified (Avin et al., 2005).

Perhaps one notable exception in which assumption (4) is more reasonable are settings in which the mediator is measured immediately after, or very shortly after, the exposure takes place. For example, Nelson et al. (1997) examine the extent to which the framing of political issues in news media affected subjects' tolerance of a Ku Klux Klan rally as mediated by affecting subjects' general political attitudes. Subjects in the study were randomized to watch one of two news clips, one of which presented the rally as a free speech issue and the other of which as a public order issue. Following the clip, survey questions were used to assess general political attitudes towards the right to free speech and the maintenance of public order and two further questions were used to assess subjects' tolerance for Klan speeches and rallies. The analysis is presented in Nelson et al. (1997) and is reanalyzed using contemporary ideas from causal inference by Imai et al. (2009). The point here, however, is that since political attitudes are assessed immediately after the exposure (the watching of the video clip) it is less likely that there is anything that is an effect of exposure that confounds the mediator-outcome relationship. Thus assumption (4) might be reasonable in this case; the exposure (video clip) was randomized so assumptions (1) and (3) would hold. Thus provided that control is adequately made for variables that confound the mediator-outcome relationship so that assumption (2) holds, one could proceed with the estimation of natural direct and indirect effects.

We will consider another similar example. In a study by Smeesters et al. (2003), the investigators used a one-trial prisoner's dilemma game with a fictitious partner to study the extent to which the effect of "morality" or "might" primes on cooperative versus competitive behavior was mediated by expectations about a partner. Expectations about the partner's behavior was assessed after exposure to either the "morality" or the "might" prime. As with the Nelson et al. (1997) study, the mediator, in this case expectations about the behavior of the partner, could be assessed immediately after the exposure and it thus may be plausible that there are no variables that are effects of exposure but confound the mediator-outcome relationship. Assumption (4) might thus be reasonable, assumptions (1) and (3) hold by randomization of the exposure and thus, provided adequate control is made for mediator-outcome confounders (assumption (2)), natural direct and indirect effects could be identified. These two examples exhibit a similar structure with the mediator being measured immediately after the exposure occurs thereby potentially rendering condition (4) more plausible. One can potentially conceive of a whole class of psychological experiments that follow a similar structure to these two examples for which assumption (4) may be more reasonable than in many other settings.

In summary, if there is an exposure-induced confounder of the mediator-outcome relationship, natural direct and indirect effects are not in general identified. Controlled direct

effects are identified in such settings (Robins and Greenland, 1992; Pearl, 2001) but require special techniques for estimation (Robins, 1999; van der Laan and Petersen, 2008; Goetgeluk et al., 2008; VanderWeele, 2009a; Vansteelandt, 2009). Unfortunately, however, controlled direct effects are of limited use in assessing mediation because the difference between a total effect and a controlled direct effect cannot in general be interpreted as an indirect effect (Kaufman et al., 2004; VanderWeele, 2009b) except in cases in which there is no interaction between the effects of the exposure and the mediator on the outcome (Robins, 2003). Natural direct and indirect effects are desirable in that they allow for effect decomposition even in settings with interaction and non-linearities but as we have seen above, the identification conditions required for these effects are in general quite stringent. The possibility of designing an experiment so that the mediator is measured or occurs immediately after the exposure, so that assumption (4) may be plausible, suggests a range of settings in which the assumptions required to identify natural direct and indirect effects may be reasonable.

CONCEPTUAL ISSUES CONCERNING INTERVENTIONS

Thus far, we have not discussed the kind of intervention or manipulation that would enable setting M to some given value m. As implied by the consistency assumption, the kind of interventions or manipulations that we consider are noninvasive in the sense that their only effect is to set the mediator at some pre-determined value m such that the intervention has no effect amongst those for whom mediator level m was naturally observed. Any conclusion that we draw in terms of controlled direct effects holds for interventions satisfying this assumption. In practice, however, it is often difficult to conceive of such noninvasive interventions, either because the mediator is not manipulable, or because any conceivable manipulation would affect the outcome in ways other through setting the mediator at some pre-determined value. Thus in the Nelson et al. (1997) study described above, the mediator was general political attitudes towards the right to free speech and the maintenance of public order. Clearly one cannot intervene on these political attitudes directly. One might conceive of some type of intervention to change these political attitudes such as the watching of another video clip; however, such interventions might well also affect the outcome in ways other than through the mediator. Different conceivable interventions to affect M so that it is set to level m might then result in different counterfactual outcomes $Y(a, m)$; the counterfactual outcome $Y(a, m)$ would then not be well defined. Clearly in many psychological experiments, similar problems will arise; when the mediator is a psychological construct, such as a particular attitude, intervening directly will not in general be possible.

In some cases, however, hypothetical interventions on the mediator may be conceivable. For example, in the study of Smeesters et al. (2003) described above, the mediator was the subjects' expectations about the behavior of the fictitious partner. One could imagine a study in which one could intervene on the subjects' expectations by having the study investigators telling the subject the behavioral decision of the fictitious partner before the subject decides on his or her own action (or alternatively telling the subject a distribution of possible actions for the fictitious partner). Once the subject is told the behavior of the partner, the expectations are effectively fixed. One thus has a potential intervention on the mediator. One still may have to be concerned with violations of the consistency assumption as to whether an individual will select the same action under a particular set of expectations as the subject would select if the subject were told the partner's behavior was that which he or she was expecting.

Further discussion of the consistency assumption and of potential violations is given in the appendix. Extending the corresponding discussion of total effects in VanderWeele (2009c), we discuss in the appendix the consistency assumptions for mediation. We clarify that essentially two things are being assumed about the interventions under consideration: first, that different conceivable interventions to fix the mediator M to some level m all give rise to the same counterfactual outcomes and, second, that interventions to set the mediator to its naturally ocurring level will give the same outcome as not intervening. The first assumption may be referred to as an assumption of *treatment-variation irrelevance* (or no-multiple-versions-of-treatment) which is needed for counterfactuals of the form $Y(a, m)$ to be well-defined and the second as an assumption of *consistency*. Both types of assumptions are subsumed by Rubin's *stable unit treatment value assumption*, abbreviated *SUTVA* (Rubin, 1990). See also VanderWeele (2009d) for discussion of potential violations of the "no-interference" component of SUTVA within a mediation context.

Even in cases in which it is not possible to conceive of a noninvasive intervention to change the mediator, the definition of a controlled direct effect can be meaningful because it effectively communicates what the exposure effect would be if nothing happened over and above fixing the mediator at some level uniformly in the population. For instance, while there are no conceivable interventions that would fix the occupation of a given subject at some pre-determined choice, and do nothing on top of that, the question of whether the father's occupation affects the respondent's income through pathways other than by influencing the respondent's occupation, is arguably a meaningful one. The fact that any conceivable intervention on the respondent's occupation would do more than just determine the occupation implies that there are no realistic interventions that would bring about exactly the controlled direct exposure effect. Nonetheless, estimates of the controlled direct exposure effect can be meaningful in this setting because they give a more pure reflection

of the direct exposure effect than could be attained through realistic interventions on the mediator.

The interpretability of natural direct and indirect effects arguably hinges to a lesser extent on the ability to control or manipulate the mediator. First, the definitions of natural direct and indirect effects only require manipulations to set the mediator to levels which are naturally occurring e.g. $M(0)$ or $M(a)$ rather than to some arbitrary level m which, for certain individuals, might never occur for any exposure a. Second, rather than conceiving of a natural direct effect such as $E[Y(1, M(0)) - Y(0, M(0))]$ as the effect of exposure intervening to fix the mediator at level $M(0)$, one might alternatively think about natural direct effects as requiring interventions that would block the exposure's effect on the mediator (Pearl, 2001; Robins, 2003). Such interventions are sometimes more easily conceivable (as well as more natural by not necessarily requiring that the mediator be uniformly fixed at the same level). Consider for instance a study on the effect of gender, A, on graduate admissions, Y, for graduate school applicants. Suppose that the goal of the study was to assess whether gender differences in admissions to a particular department are entirely explained by the admission committee's perception of gender, M, and thus potential discrimination based on gender. In principle, blinding of gender at the time of application (e.g. having all applicants in the study list either "male" as gender, or having all applicants in the study list "female" as gender) would bring about the natural direct effect of gender on admission because letting say $A = 1$ denote female and $A = 0$ denote male, we would have that $E[Y(1, M(0)) - Y(0, M(0))] = E[Y(1, 0) - Y(0, 0)]$. In this case the natural direct effect and the controlled direct effect in fact coincide. Robins (2003) notes that generally natural direct effects can be only interpreted as the effect of exposure on the outcome intervening to block the effect of the exposure on the mediator if the intervention were to block the first conceivable link on the pathway from exposure to mediator.

The issue of the applicability of hypothetical interventions arise in causal inference outside the mediation context, when one is only examining total effects (Hernán, 2005; van der Laan et al., 2005). In the occupation example, it is clearly as inconceivable to fix the occupation of a respondent's father (the exposure A) at some pre-determined choice, and do nothing on top of that, as it is to do so for that of the respondent (the mediator M). Many exposures or causes of interest are not obviously manipulable; examples of such exposures might include gender, race or psychological disposition. Without clear interventions to change these variables, counterfactual contrasts are not necessarily well-defined or are ill-defined to the extent to which the intervention is not specified (Lewis, 1973; Robins and Greenland, 2000). However, in general we still are interested in examining the effects of these variables. The counterfactual or potential outcomes framework is perhaps not ideally suited, at least in its present form, to conceptualize the effects of such non-manipulable exposures. Nevertheless, as discussed above, the consistency assumption suggests that the inferred effects can perhaps be interpreted as what would be realized under non-invasive interventions. Some further progress can perhaps be made by relaxing or refining the consistency assumption and allowing for multiple versions of treatment (van der Laan et al., 2005; Taubman et al., 2008; VanderWeele, 2009c). However, an alternative approach, which perhaps more closely ties causation to physical laws governing systems, may be desirable (Commenges and Gegout-Petit, 2009). Work has been done in causal inference on direct and indirect effects outside of the counterfactual framework (Didelez et al., 2007; Geneletti, 2008) but this work also conceptualizes direct and indirect effects through various potential interventions. The point here, however, is that the issue of the manipulability of variables is not unique to the context of mediation but arises even in the context of examining total effects.

We conclude this section by noting that in cases in which interventions on the exposure of interest are conceivable but interventions on the mediator are not, an alternative approach to assessing mediation is possible using concepts of principal stratification (Frangakis and Rubin, 2002; Rubin, 2004; VanderWeele, 2008; Gallop et al., 2009). Principal strata are strata defined by the joint counterfactual outcomes $M(a)$ for all possible values of a. One might assess direct effects, for example, by examining the effect of exposure on the outcome for individuals for whom the exposure does not affect the value of the mediator e.g. for whom $M(0) = M(a)$. Such an approach is, in a certain sense, advantageous in that it only requires counterfactuals $M(a)$ and $Y(a)$, rather than also counterfactuals of the form $Y(a, m)$ i.e. it does not require hypothetical interventions on the mediator. Unfortunately, however, the utility of the approach based on principal stratification is limited because of the inability to identify which individuals fall into which principal strata, because the probability of falling within the principal stratum $M(a) = M(0)$ will be zero in many realistic applications (Robins et al., 2007), and because the notions of mediation based on principal direct effects do not correspond in any clear way to mechanisms of scientific interest (Joffe et al., 2007).

CONCEPTUAL ISSUES CONCERNING COMPOSITION

In this section we would like to discuss briefly the assumption that is sometimes referred to as composition (Pearl, 2000) that $Y(a) = Y(a, M(a))$ i.e. that the potential outcome $Y(a)$ intervening to set A to a is equal to the potential outcome $Y(a, M(a))$ intervening to set A to a and to set M to the value it would have been if A had been a. This assumption allows one to express the natural direct effect, $E[Y(1, M(0)) - Y(0, M(0))|C]$, as $E[Y(1, M(0)) - Y(0)|C]$

and allows one to decompose a total effect into a natural direct and indirect effect:

$$E[Y(a)-Y(0)|C]=E[Y(a,M(a))-Y(a,M(0))|C] + E[Y(a,M(0))-Y(0,M(0))|C].$$

The composition assumption contrasts with the consistency assumption for $Y(a,m)$ which states that when $A=a$ and $M=m$ then $Y(a,m)=Y$ i.e. that amongst the subgroup with observed exposure $A=a$ and observed mediator $M=m$, the observed outcome Y is equal to the value of Y that would have been obtained intervening to set A to a and M to m.

Because the composition assumption is used for effect decomposition, it is an important assumption in the context of mediation. The assumption is, however, often not explicitly stated but merely implicitly assumed. Perhaps this is in part because the composition assumption arguably does not involve substantial conceptual issues above and beyond those entailed by the consistency and treatment-variation irrelevance assumptions. The consistency and treatment-variation irrelevance assumptions effectively presuppose hypothetical interventions on the exposure and the mediator; the composition assumption also presupposes hypothetical interventions on the exposure and the mediator but only imposes restrictions concerning hypothetical interventions on the mediator for levels which naturally occur under various interventions on the exposure. The consistency assumption requires that when $A=a$ and $M=m$ then $Y(a,m)=Y$ and thus that the value of Y under two interventions to set A and M to their natural levels simply equals the observed outcome; the composition assumption requires that $Y(a)=Y(a,M(a))$ and thus that, under interventions on a, interventions on M to set it to its naturally occurring level $M(a)$ have no further effect on the outcome obtained. The consistency and treatment-variation irrelevance assumptions thus essentially presuppose that interventions on both A and M are in some sense noninvasive, as discussed above and also in the appendix, while the composition assumption essentially just presupposes that interventions on M are noninvasive. The conceptual issues involved in the composition assumption thus do not seem to entail considerably more than that which was required with the consistency assumption. The consistency assumption does not mathematically entail the composition assumption but we find it difficult to imagine cases in which a researcher would be willing to make the consistency assumption but unwilling to make the composition assumption.

CONCLUDING REMARKS

In this paper we have related concepts of direct and indirect effects from causal inference to concepts arising from the approach of Baron and Kenny (1986), popular in the social sciences. We have also considered a number of important conceptual issues concerning mediation. We have seen that in certain psychological experiments the assumptions required to identify natural direct and indirect effects may be rendered more plausible. This is important because it is notions of natural direct and indirect effects which allow for effect decomposition even in settings involving interactions and non-linearities. We have also considered at length the issues of the relation of hypothetical interventions on the mediator and the exposure to the counterfactual framework. We have seen that often it is difficult to conceive of interventions that non-invasively fix the mediator and have discussed potential violations of the so-called consistency assumption. We have furthermore seen that in some cases interventions on the mediator are conceivable and that, moreover, for natural direct and indirect effects, one only need conceive of interventions on the mediator to set the mediator to naturally occurring levels. In addition, natural direct and indirect effects can in some cases be conceived of as effects that would result by blocking the effect of an exposure on the mediator.

Nevertheless, the counterfactual framework is, at least at present, tied quite closely to the notion of a hypothetical intervention and in some cases this can be problematic, even when total effects are in view. The formalization of the counterfactual approach is still arguably of use even in settings in which conceiving of interventions is difficult. Questions of total effects and of mediation are likely to be of interest in practice even when interventions on the considered exposures and mediators are not possible; the counterfactual approach clarifies at least the assumptions and identification conditions required for assessing direct and indirect effects and makes clear when violations of these assumptions will lead to bias. The counterfactual framework is a simplification of a complex reality but it arguably moves us one step closer to the quantities of interest when we think about mediation.

Finally, we have extended our discussion of interventions in the mediation context to the composition assumption, an assumption that is not always explicitly stated but is often utilized when natural direct and indirect effects are in view. We have seen that when the various consistency and treatment-variation irrelevance assumptions hold, the composition assumption does not impose considerable conceptual challenges above and beyond those already implicit in the consistency assumptions. The notions of mediation using the counterfactual approach are sometimes subtle, but we hope that the contributions in this paper will help clarify the conceptual issues involved in utilizing the counterfactual framework to address these questions of mediation in the social sciences and beyond.

ACKNOWLEDGEMENTS

We would like to thank the participants at a session on "Mediation and Causal Inference" at the 2009 ENAR International Biometric Society meetings for the interesting discussion at that session that prompted the writing of this paper. We also thank Haiqun Lin for the invitation to write

the paper and we thank an anonymous referee for helpful comments.

APPENDIX

Controlled and natural direct and indirect effects using regression

If the regression models (6) and (8) are correctly specified and assumptions (1) and (2) hold then we could compute the controlled direct effect as follows:

$$\begin{aligned}
&E[Y(a,m)-(a^*,m)|C=c]\\
&\quad=E[Y|C=c,A=a,M=m]\\
&\qquad-E[Y|C=c,A=a^*,M=m]\\
&\quad=(\theta_0+\theta_1a+\theta_2m+\theta_3am+\theta_4'c)\\
&\qquad-(\theta_0+\theta_1a^*+\theta_2m+\theta_3a^*m+\theta_4'c)\\
&\quad=(\theta_1a+\theta_3am-\theta_1a^*-\theta_3a^*m)\\
&\quad=\theta_1(a-a^*)+\theta_3m(a-a^*)
\end{aligned}$$

If the regression models (6) and (8) are correctly specified and assumptions (1)–(4) hold, we could compute natural direct effects by

$$\begin{aligned}
&E[Y(a,M(a^*))-Y(a^*,M(a^*))|C=c]\\
&\quad=\sum\nolimits_m\{E[Y|C=c,A=a,M=m]\\
&\qquad-E[Y|C=c,A=a^*,M=m]\}\\
&\qquad\times P(M=m|C=c,A=a^*)\\
&\quad=\sum\nolimits_m\{(\theta_0+\theta_1a+\theta_2m+\theta_3am+\theta_4'c)\\
&\qquad-(\theta_0+\theta_1a^*+\theta_2m+\theta_3a^*m+\theta_4'c)\}\\
&\qquad\times P(M=m|C=c,A=a^*)\\
&\quad=\sum\nolimits_m\{(\theta_1a+\theta_2m+\theta_3am)\\
&\qquad-(\theta_1a^*+\theta_2m+\theta_3a^*m)\}\\
&\qquad\times P(M=m|C=c,A=a^*)\\
&\quad=\{\theta_1a+\theta_2E[M|A=a^*,C=c]\\
&\qquad+\theta_3aE[M|A=a^*,C=c])\\
&\qquad-(\theta_1a^*+\theta_2E[M|A=a^*,C=c]\\
&\qquad+\theta_3a^*E[M|A=a^*,C=c])\}\\
&\quad=\{\theta_1a+\theta_2(\beta_0+\beta_1a^*+\beta_2'c)\\
&\qquad+\theta_3a(\beta_0+\beta_1a^*+\beta_2'c))\\
&\qquad-(\theta_1a^*+\theta_2(\beta_0+\beta_1a^*+\beta_2'c)\\
&\qquad+\theta_3a^*(\beta_0+\beta_1a^*+\beta_2'c))\}\\
&\quad=\{\theta_1a+\theta_3a(\beta_0+\beta_1a^*+\beta_2'c)\\
&\qquad-(\theta_1a^*+\theta_3a^*(\beta_0+\beta_1a^*+\beta_2'c))\}\\
&\quad=(\theta_1+\theta_3\beta_0+\theta_3\beta_1a^*+\theta_3\beta_2'c)(a-a^*)
\end{aligned}$$

If the regression models (6) and (8) are correctly specified and assumptions (1)–(4) hold, we could compute natural indirect effects by

$$\begin{aligned}
&E[Y(a,M(a))-Y(a,M(a^*))|C=c]\\
&\quad=\sum\nolimits_m E[Y|C=c,A=a,M=m]\\
&\qquad\times P(M=m|C=c,A=a)\\
&\qquad-\sum\nolimits_m E[Y|C=c,A=a,M=m]\\
&\qquad\times P(M=m|C=c,A=a^*)\\
&\quad=\sum\nolimits_m(\theta_0+\theta_1a+\theta_2m+\theta_3am+\theta_4'c)\\
&\qquad\times P(M=m|C=c,A=a)\\
&\qquad-\sum\nolimits_m(\theta_0+\theta_1a+\theta_2m+\theta_3am+\theta_4'c)\\
&\qquad\times P(M=m|C=c,A=a^*)\\
&\quad=(\theta_0+\theta_1a+\theta_2E[M|A=a,C=c]\\
&\qquad+\theta_3aE[M|A=a,C=c]+\theta_4'c)\\
&\qquad-\sum\nolimits_c(\theta_0+\theta_1a+\theta_2E[M|A=a^*,C=c]\\
&\qquad+\theta_3aE[M|A=a^*,C=c]+\theta_4'c)\\
&\quad=(\theta_0+\theta_1a+\theta_2(\beta_0+\beta_1a+\beta_2'c)\\
&\qquad+\theta_3a(\beta_0+\beta_1a+\beta_2'c)+\theta_4'c)\\
&\qquad-(\theta_0+\theta_1a+\theta_2(\beta_0+\beta_1a^*+\beta_2'c)\\
&\qquad+\theta_3a(\beta_0+\beta_1a^*+\beta_2'c)+\theta_4'c)\\
&\quad=\theta_2\beta_1(a-a^*)+\theta_3\beta_1a(a-a^*).
\end{aligned}$$

Standard errors of controlled and natural direct and indirect effects using regression

Suppose that models (6) and (8) have been fit using standard linear regression software and that the resulting estimates $\hat{\beta}$ of $\beta\equiv(\beta_0,\beta_1,\beta_2')'$ and $\hat{\theta}$ of $\theta\equiv(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4')'$ have covariance matrices Σ_β and Σ_θ, which can be obtained from most off-the-shelf statistical software packages. Then the covariance matrix of $(\hat{\beta}',\hat{\theta}')'$ is

$$\Sigma\equiv\begin{pmatrix}\Sigma_\beta & 0\\ 0 & \Sigma_\theta\end{pmatrix},$$

which can be seen upon noting that

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta},\hat{\theta})&=E\big\{\mathrm{Cov}(\hat{\beta},\hat{\theta}|M,A,C)\big\}\\
&\quad+\mathrm{Cov}\big\{E(\hat{\beta}|M,A,C),E(\hat{\theta}|M,A,C)\big\}\\
&=0+\mathrm{Cov}\big(\hat{\beta},\theta\big)=0,
\end{aligned}$$

where we use the fact that $\hat{\beta}$ is a function of M,A and C only. Standard errors of the controlled and natural direct and indirect effects in (9) can then be obtained (using the Delta method) as

$$\sqrt{\Gamma\Sigma\Gamma'}|a-a^*|$$

with $\Gamma\equiv(0,0,0',0,1,0,m,0')$ for the controlled direct effect in (9), $\Gamma\equiv(\theta_3,\theta_3a^*,\theta_3C',0,1,0,\beta_0+\beta_1a^*+\beta_2'C,0')$ for the pure natural direct effect in (9) (the same expression holds for the total natural direct effect upon substituting a and a^*)

and $\Gamma \equiv (0, \theta_2 + \theta_3 a, 0', 0, 0, \beta_1, \beta_1 a, 0')$ for the total natural indirect effect in (9) (the same expression holds for the pure natural indirect effect upon substituting a and a^*). In these expressions, $0'$ denotes a row vector of the dimension of C, containing zeroes only.

Treatment-variation irrelevance assumptions and consistency assumptions for mediation

In this appendix, we extend the discussion of the consistency and treatment-variation irrelevance assumptions for total effects in VanderWeele (2009c) to the case of mediation and potential interventions on a mediator. For each possible exposure level a, let $\mathcal{K}_a$ denote some set of interventions to fix exposure to level a. For $k_a \in \mathcal{K}_a$, let $M_j(a, k_a)$ and $Y_j(a, k_a)$ be the values of the mediator and the outcome respectively that would be observed for individual j under an intervention to fix exposure to level a by means k_a. As in VanderWeele (2009c), we consider two distinct assumptions. First, if for each a we have,

(C1) $\quad Y_j(a, k_a) = Y_j(a, k'_a)$ for all $k_a, k'_a \in \mathcal{K}_a$

then we say that the assumption of treatment-variation irrelevance holds for Y with respect to A and the potential outcome $Y_j(a)$ can be defined as $Y_j(a) := Y_j(a, k_a)$ for any $k_a \in \mathcal{K}_a$. Second, if assumption (C1) holds then we say that the consistency assumption for $Y(a)$ is satisfied if for each j,

(C2) $\quad$ for some $k_a \in \mathcal{K}_a$, $Y_j = Y_j(a, k_a)$ when $a = A_j$.

Assumption (C1) requires for each a that the potential outcomes $Y_j(a, k_a)$ take the same value irrespective of what means k_a is used to set A to a so long as $k_a \in \mathcal{K}_a$. Assumption (C2) then requires that for some k_a the potential outcome $Y_j(a, k_a)$ is equal to the observed outcome Y_j when $a = A_j$. Assumption (C1) captures the notion that the set of interventions under consideration, $\mathcal{K}_a$, do not affect Y except through setting A to level a. Assumption (C2) captures the notion that the interventions under consideration are non-invasive in that the outcome that would be observed under an intervention to set exposure to the level it actually was is equal to the outcome that was in fact observed (i.e. the naturally occurring outcome).

We can formulate similar assumptions for $M(a)$. If for each a,

(C3) $\quad M_j(a, k_a) = M_j(a, k'_a)$ for all $k_a, k'_a \in \mathcal{K}_a$

then we say that the assumption of treatment-variation irrelevance holds for M with respect to A and then the potential outcome $M_j(a)$ can be defined as $M_j(a) := M_j(a, k_a)$ for any $k_a \in \mathcal{K}_a$. If assumption (C3) holds then we say that the consistency assumption for $M(a)$ is satisfied if for each j,

(C4) $\quad$ for some $k_a \in \mathcal{K}_a$, $M_j = M_j(a, k_a)$ when $a = A_j$.

We can furthermore consider similar assumptions for $Y(a, m)$. For each possible mediator level m, let $\mathcal{K}_m$ denote some set of interventions to fix the mediator to level m. For $k_a \in \mathcal{K}_a$ and $k_m \in \mathcal{K}_m$, let $Y_j(a, m, k_a, k_m)$ be the value of the outcome for individual j that would be observed under interventions to fix exposure to level a by means k_a and to fix the mediator to level m by means k_m. If for each a and m,

(C5) $\quad Y_j(a, m, k_a, k_m) = Y_j(a, m, k'_a, k'_m)$ for all $k_a, k'_a \in \mathcal{K}_a$, $k_m, k'_m \in \mathcal{K}_m$

then we say that the assumption of treatment-variation irrelevance holds for Y with respect to (A, M) and the potential outcome $Y_j(a, m)$ can be defined as $Y_j(a, m) := Y_j(a, m, k_a, k_m)$ for any $k_a \in \mathcal{K}_a$ and $k_m \in \mathcal{K}_m$. If assumption (C5) holds then we say that the consistency assumption for $Y(a, m)$ is satisfied if for each j,

(C6) $\quad$ for some $k_a \in \mathcal{K}_a$ and some $k_m \in \mathcal{K}_m$, $Y_j = Y_j(a, m, k_a, k_m)$ when $a = A_j$ and $m = M_j$.

VanderWeele (2009c) noted that treatment-variation irrelevance assumptions, such as (C1), (C3) and (C5), are necessary not only in order to articulate the consistency assumptions but also in order to articulate no-unmeasured-confounding assumptions and even for the potential outcomes of the form $Y_j(a)$, $M_j(a)$ and $Y_j(a, m)$ to be well-defined. Similarly, these treatment-variation irrelevance assumptions are necessary in order to articulate the composition assumption as $Y(a) = Y(a, M(a))$.

The treatment-variation irrelevance and consistency assumptions given above can, however, potentially be made more plausible by employing stochastic counterfactuals (Robins and Greenland, 1989) and allowing the counterfactuals, $M_j(a, k_a), Y_j(a, k_a), Y_j(a, m, k_a, k_m)$, and the actual outcomes, M_j, Y_j, to follow a distribution for each j, rather than simply being single values. See VanderWeele (2009c) for further discussion.

Received 4 June 2009

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.

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometics*, **64**, 695–701.

Cole, S.R. and Hernán, M.A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, **31**, 163–165.
Commenges, D. and Gegout-Petit, A. (2009). A general dynamical statistical model with possible causal interpretation. *Journal of the Royal Statistical Society, Series B*, **71**, 719–736.
Didelez, V., Dawid, A. P. and Geneletti, S. (2006) Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*.
Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
Gallop, R., Small, D.S., Lin, J.Y., Elliott, M.R., Joffe, M. and Ten Have, T.R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, **28**, 1108–1130.
Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B*, **69**, 199–216.
Goetgeluk, S., Vansteelandt, S. and Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B*, **70**, 1049–1066.
Hafeman, D.M. and VanderWeele, T.J. (2009). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, in press.
Hernán, M.A. (2004). A definition of causal effect for epidemiological studies. *Journal of Epidemiology and Community Health*, **58**, 265–271.
Hernán, M.A. (2005). Invited commentary: Hypothetical interventions to define causal effects: afterthought or prerequisite? *American Journal of Epidemiology*, **162**, 618–620.
Imai, K., Keele, L. and Yamamoto, T. (2009). Identification, inference, and sensitivity analysis for causal mediation effects. Working paper. Web address: http://imai.princeton.edu/research/files/mediation.pdf.
Joffe, M., Small, D. and Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*, **22**, 74–97.
Judd, C.M. and Kenny, D.A. (1981). Process analysis: estimating mediation in treatment evaluations. *Evaluation Review*, **5**, 602–619.
Kaufman, J.S., MacLehose, R.F. and Kaufman, S. (2004). A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives and Innovations*, **1**, 4.
Kaufman, S., Kaufman, J.S., MacLehose, R.F., Greenland, S. and Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine*, **24**, 1683–1702.
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*, **139**, 3473-3487.
Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge.
MacKinnon, D.P. (2008). *An Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York.
Nelson, T.E., Clawson, R.A., and Oxley, Z.M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, **91**, 567–583.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.
Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. San Francisco: Morgan Kaufmann, 411–420.
Petersen, M.L., Sinisi, S.E. and van der Laan, M.J. (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: Glymour C., Cooper G.F., eds. *Computation, Causation, and Discovery*. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, 349–405.
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, Oxford University Press, New York, 70–81.
Robins, J.M. and Greenland, S. (1989) The probability of causation under a stochastic model for individual risk. *Biometrics*, **45**, 1125–1138.
Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
Robins, J.M. and Greenland, S. (2000). Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid. *Journal of the American Statistical Association*, **95**, 477–482.
Robins, J.M., Rotnitzky, A. and Vansteelandt, S. (2007). Discussion of 'Principal stratification designs to estimate input data missing due to death' by C.E. Frangakis, D.B. Rubin, M.-W. An, and E. MacKenzie. *Biometrics*, 63, 650–653.
Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, **6**, 34–58.
Rubin, D.B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, **25**, 279–292.
Rubin, D.B. (2004). Direct and indirect effects via potential outcomes. *Scandinavian Journal of Statistics*, **31**, 161–170.
Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, **28**, 558–571.
Smeesters, D., Warlop, L., Van Avermaet, E., Corneille, O. and Yzerbyt, V.Y. (2003). Do not prime hawks with doves: the interplay of construct activation and consistency of social value orientation on cooperative behavior. *Journal of Personality and Social Psychology*, **84**, 972–987.
Taubman, S.L., Robins, J.M., Mittleman, M.A. and Hernán, M.A. (2008). Alternative approaches to estimating the effects of hypothetical interventions. In *Joint Statistical Meetings Proceedings*. Alexandria, VA: American Statistical Association.
van der Laan, M.J., Haight, T.J. and Tager, I.B. (2005). Response to: "Hypothetical interventions to define causal effects" by M.A. Hernán. *American Journal of Epidemiology*, **162**, 621–622.
van der Laan, M.J. and Petersen, M.L. (2008). Direct effect models. *International Journal of Biostatistics*, **4**: Article 23.
VanderWeele, T.J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78:2957–2962.
VanderWeele, T.J. (2009a). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, **20**, 18–26.
VanderWeele, T.J. (2009b). Mediation and mechanism. *European Journal of Epidemiology*, **24**, 217–224.
VanderWeele, T.J. (2009c). Concerning the consistency assumption in causal inference. *Epidemiology*, **20**, 880-883.
VanderWeele, T.J. (2009d). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods and Research*, in press.
Vansteelandt, S. (2009). Estimating direct effects in cohort and case-control studies. *Epidemiology*, **20**, 851-860.

Tyler J. VanderWeele
Departments of Epidemiology and Biostatistics,
Harvard University, 677 Huntington Avenue,
Boston, MA, 02115, U.S.A.
E-mail address: tvanderw@hsph.harvard.edu

Stijn Vansteelandt
Department of Applied Mathematics
and Computer Sciences, Ghent University
281 (S9) Krijgslaan, 9000 Ghent, Belgium
E-mail address: Stijn.Vansteelandt@ugent.be