Published in final edited form as:

Epidemiology. 2012 May ; 23(3): 454–463. doi:10.1097/EDE.0b013e31824d5fe7.

Mediation analysis with multiple versions of the mediator

Tyler J. VanderWeele

Departments of Epidemiology and Biostatistics, Harvard School of Public Health

Abstract

The causal inference literature has provided definitions of direct and indirect effects based on counterfactuals that generalize the approach found in the social science literature. However, these definitions presuppose well defined hypothetical interventions on the mediator. In many settings there may be multiple ways to fix the mediator to a particular value and these different hypothetical interventions may have very different implications for the outcome of interest. In this paper we consider mediation analysis when multiple versions of the mediator are present. Specifically, we consider the problem of attempting to decompose a total effect of an exposure on an outcome into the portion through the intermediate and the portion through other pathways. We consider the setting in which there are multiple versions of the mediator but the investigator only has access to data on the particular measurement, not which version of the mediator may have brought that value about. We show that the quantity that is estimated as a natural indirect effect using only the available data does indeed have an interpretation as a particular type of mediated effect; however, the quantity estimated as a natural direct effect in fact captures both a true direct effect and an effect of the exposure on the outcome mediated through the effect of the version of the mediator that is not captured by the mediator measurement. The results are illustrated using two examples from the literature, one in which the versions of the mediator are unknown and another in which the mediator itself has been dichotomized.

Introduction

The causal inference literature has provided theory and methods to estimate direct and indirect effects^{1–11}, or in other words, to assess mediation, the extent to which an exposure affects an outcome through a particular intermediate versus through other pathways. This literature in causal inference generalizes the existing methodological literature in the social sciences^{12–14} by allowing for settings with interactions and non-linearities and by clarifying the no-unmeasured-confounding assumptions that are required for the estimation of such effects. The definitions of direct and indirect effects provided by the causal inference literature make use of potential or counterfactual outcomes¹⁵, generally conceived of as the outcome that would have arisen under various hypothetical, possibly contrary to fact, interventions. In the context of mediation, interventions are conceived of both on the exposure and on the mediator. It is sometime stated that such potential or counterfactual outcomes are well-defined only to the extent that the investigator has clearly specified the intervention in view^{16,17}. Moreover, one's estimates of causal effects will only reasonably correspond to the intervention envisioned if the exposure that naturally occurred and an intervention to set the exposure to this same level would give rise to the same outcome; in the technical language of the causal inference literature this is sometimes referred to as the “consistency” assumption^{17–19}. These assumptions are problematic in settings in which there are multiple versions of the treatment or exposure; or, within the context of mediation,

when there are multiple ways to set the mediator to a particular value if these different hypothetical interventions to fix the mediator to the same value may have very different consequences for the outcome.

For example, suppose the treatment were an educational intervention, the mediator a measure of classroom quality and the outcome some measure of average test scores. In this context, there may be multiple ways to go about changing classroom quality (giving the teacher additional coaching, replacing the teacher, providing a more adequate room for the class etc.). Even if each of these interventions on the mediator resulted in the same level of classroom quality under some measure, they may have very different consequences for the test scores of the class. Conceived of another way, we have “multiple versions of the mediator.” This creates difficulties for the definition and interpretation of direct and indirect effects in the counterfactual framework. It is not clear what is necessarily meant by mediation when there are many ways the mediator may take on a particular value and when these have very different effects on the outcome. As another example, Caffo et al.²⁰ studied the extent to which the effects of cumulative lead exposure in organolead manufacturers on cognitive function scores was mediated by brain volumes. Here, with the mediator, brain volumes, it is not really clear how to intervene on this variable at all or what set of potential hypothetical interventions we may be considering. In these settings, multiple versions of the mediator create difficulties for the counterfactual-based approach to mediation analysis.

An approach has, however, begun to develop within the causal inference literature to facilitate thinking about inference in settings with multiple versions of treatment.^{19,21,22} In this paper we extend and apply this approach to the question of mediation when there are multiple versions of the mediator. We will consider two settings: one in which there may be multiple ways to intervene on the variable taken as the mediator in the analysis and one in which the actual mediator has been coarsened (e.g. dichotomized) so that one value of the mediator measure corresponds to multiple values of the true mediator. The mathematics behind these two settings are essentially identical and so we will be able to consider both of them together. We will not consider the setting, addressed in some of the social science literature, in which multiple measures of an underlying latent mediator are available for each person in the study.

In what follows, we will call the underlying true mediator or the set of interventions to fix the mediator value the “version” and the variable actually used in the analysis the “mediator measure”. We will assume investigator does not have access to data on “version” but proceeds with the estimation of direct and indirect effects with the data on the mediator measure available (e.g. classroom quality or brain volumes in the examples above) and we will consider what the interpretation of the direct and indirect effect estimates that result. We will show that these effects can be interpreted as the effects that would have arisen had the version of the mediator been randomly selected from the distribution of the versions that arise naturally in certain subpopulations. In particular, we show that what is estimated as an indirect effect does in fact correspond to a particular mediated effect wherein the version of the mediator is randomly selected from the exposed with particular values of the measured mediator; however, we also show that what is estimated as a direct effect is in fact equal to the sum of a direct effect, plus a mediated effect corresponding to the effect of the exposure on the outcome mediated by versions of the mediator that is not captured by the mediator measure itself. We will consider the implications of these results by illustration to a couple of existing mediation analyses in the literature.

Direct and Indirect Effects: Notation and Definitions

In this section we will review the definitions, assumptions and estimation results from counterfactual approach to mediation when there is a single well-defined intervention on the mediator. In the following section, we will consider how this approach is altered when there are multiple versions of the mediator. Here, we let A denote the exposure of interest, Y the outcome of interest, and M the potential mediator of interest. For example, A might denote cumulative lead exposure, Y might denote cognitive functioning test scores, and M might denote brain volumes. We will let C denote a set of baseline covariates not affected by the exposure. We will let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a . We will let Y_{am} denote the value of the outcome that would have been observed had the exposure, A , been set to level a , and had the mediator, M , been set to level m . These counterfactual or potential outcome variables, Y_a , M_a and Y_{am} all presuppose that at least hypothetical interventions on A and M are conceivable. A further assumption is often generally made, sometimes referred to as the consistency assumption, that when $A = a$, the counterfactual outcomes Y_a and M_a are, respectively, equal to the observed outcomes Y and M . We likewise assume that when $A = a$ and $M = m$, the counterfactual outcome Y_{am} is equal to Y .⁶ This assumption will be relaxed somewhat in the following section.

The average controlled direct effect, conditional on covariates $C = c$, comparing exposure level $A = 1$ to $A = 0$ and fixing the mediator to level m is defined by $E[Y_{1m} - Y_{0m}|c]$ and captures the effect of exposure A on outcome Y , intervening to fix M to m ; it may be different for different levels of m .^{1,2} The direct effect is always relative to the mediator M being considered (i.e. through pathways other than M). The natural direct effect,^{1,2} conditional on covariates $C = c$, is defined as $E[Y_{1M_0} - Y_{0M_0}|c]$ and differs from controlled direct effects in that the intermediate M is set to the level M_0 , the level that it would have naturally been under some reference condition for the exposure, $A = 0$. Similarly, the average natural indirect effect, conditional on $C = c$, can be defined as $E[Y_{1M_1} - Y_{1M_0}|c]$, which compares the effect of the mediator at levels M_1 and M_0 on the outcome when exposure is set to $A = 1$. For the natural indirect effect to be non-zero, the exposure would have to change the mediator and that change in the mediator would have to change the outcome: natural indirect effects thus capture formally our notion of mediation. Natural direct and indirect effects are referred to by Robins and Greenland¹ as “pure direct effects” and “total indirect effects” respectively. Natural direct and indirect effects have the property that a total effect, $E[Y_1 - Y_0|c]$, decomposes into a natural direct and indirect effect: $E[Y_1 - Y_0|c] = E[Y_{1M_1} - Y_{0M_0}|c] = E[Y_{1M_1} - Y_{1M_0}|c] + E[Y_{1M_0} - Y_{0M_0}|c]$; the decomposition holds even when there are interactions and non-linearities. The approach of Baron and Kenny¹³ to direct and indirect effects, common in the social sciences, does not in general allow for effect decomposition in the presence of interactions and non-linearities.

Direct and indirect effects in general require stronger no-unmeasured confounding assumptions than total effects to estimate. The following four assumptions suffice to identify natural direct and indirect effects from data^{2,6}: (i) the effect the exposure A on the outcome Y is unconfounded conditional on C ; (ii) the effect the mediator M on the outcome Y is unconfounded conditional on C ; (iii) the effect the exposure A on the mediator M is unconfounded conditional on C ; and (iv) there is no effect of the exposure that itself confounds the mediator-outcome relationship. Only assumptions (i) and (ii) are required to estimate controlled direct effects. More formal articulations of these assumptions in terms of counterfactual independence are given in the appendix. Under these four no-unmeasured confounding assumptions, we can estimate direct and indirect effects from observed data. The natural indirect effect, conditional on covariates $C = c$, is equal to

$$Q_1 := \sum_m E[Y|A=1, m, c] \{P(m|A=1, c) - P(m|A=0, c)\}.$$

i.e. the conditional expectation of the outcome, standardized by the distribution of the intermediate when $A = 1$ versus when $A = 0$. The natural direct effect, conditional on $C = c$, is equal to:

$$Q_2 := \sum_m \{E[Y|A=1, m, c] - E[Y|A=0, m, c]\} P(m|A=0, c)$$

i.e. average difference in outcomes comparing $A = 1$ to $A = 0$, conditional on $M = m$, $C = c$, standardized by the distribution of the intermediate when $A = 0$. Finally, the controlled direct effect with the intermediate fixed to level m , conditional on covariates $C = c$, is simply equal to:

$$Q_3 := E[Y|A=1, m, c] - E[Y|A=0, m, c]$$

i.e. the average difference in outcomes comparing $A = 1$ to $A = 0$, conditional on $M = m$, $C = c$. In the following section we will consider the interpretation of these three quantities, Q_1 , Q_2 , Q_3 , when there are in fact multiple versions of the mediator.

In the setting in which there are not multiple versions of the mediator, however, the quantities above are equal to the natural indirect, natural direct and controlled direct effects respectively, provided again that assumptions (i)–(iv) hold. There are then a variety of ways to go about such estimation in practice. One particularly straightforward method involves using a regression model for the outcome Y and combining this with the results from a regression model for M . This approach essentially generalizes that of Baron and Kenny sometimes also called the “product method” in the social sciences literature to allow for potential exposure-mediator interaction. Thus, for example, VanderWeele and Vansteelandt⁶ showed that if assumptions (i)–(iv) hold and if Y and M are continuous and the following regression models for Y and M are correctly specified:

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

then the natural indirect and direct effects, and the controlled direct effect, conditional on covariates $C = c$, are given respectively by

$$\begin{aligned} E[Y_{1M_1} - Y_{1M_0} | c] &= \theta_2 \beta_1 + \theta_3 \beta_1 a \\ E[Y_{1M_0} - Y_{0M_0} | c] &= (\theta_1 + \theta_3 \beta_0 + \theta_3 \beta_2' c) \quad (1) \\ E[Y_{1m} - Y_{0m} | c] &= \theta_1 + \theta_3 m. \end{aligned}$$

If there is no interaction between A and M so that $\theta_3 = 0$, then these expressions reduce to the expressions often employed in the psychology literature: the controlled direct effect and the natural direct effect are then both equal to θ_1 and the natural indirect effect is $\theta_2 \beta_1$. More general expressions for non-binary treatment are given by VanderWeele and Vansteelandt⁶, along with standard errors for these estimators. VanderWeele and Vansteelandt⁷ also give other expressions for a binary outcome Y ; and Valeri and VanderWeele²³ give similar formulae when the mediator is binary. A more general approach, based on simulations,

rather than analytic formulae, is given by Imai et al.⁹ Sensitivity analysis techniques are also available when the no-unmeasured-confounding assumptions are violated to assess the extent to which conclusions and estimates might be altered by confounding variables for which control has not been made.^{7,9,11}

Mediation Analysis with Multiple Versions of the Mediator

Suppose now that for every possible value of the mediator, $M = m$, there are potentially multiple conceivable interventions which would allow M to be fixed at level m . We will let K denote the underlying version of the mediator. Several different values of K may correspond to the actual value of M (the variable M is effectively a coarsening of K). Suppose now then that K is the underlying version of the mediator, but we only observe M , our measurement of the mediator, and proceed with estimating direct and indirect effects using M . A causal diagram illustrating the relationships is given in Figure 1. What then is the interpretation of the direct and indirect effects that we estimate using data only on M ? The next several results provide an answer to this question. Further technical details are given in the appendix.

In the results that follow we will assume that the covariates C suffice for the analogues of no-unmeasured-confounding assumptions (i)–(iv) to hold with respect to the effects of A and K on Y . That is, we will assume: (i*) the effect the exposure A on the outcome Y is unconfounded conditional on C ; (ii*) the effect the version K on the outcome Y is unconfounded conditional on C ; (iii*) the effect the exposure A on the version K is unconfounded conditional on C ; and (iv*) there is no effect of the exposure that itself confounds the version-outcome relationship. Formal statements of these in terms of counterfactual independence are given in the appendix. We note that it may be difficult to satisfy these no-unmeasured confounding assumptions in practice if we do not know what precisely are the version of the mediator that may be present in a population, a point we return to again below.

Before stating the results it will be helpful to have a bit more additional notation. Let Y_{ak} be the outcome that would have occurred if exposure had been set to level a and the version of the mediator to level k . Let G_a be a random draw from the distribution of the version of the mediator that involves first randomly selecting a value of M from amongst those with $A = a$, $C = c$ and then randomly selecting a version of the mediator from amongst those with $M = m$, $A = 1$, $C = c$. We will then let Y_{1G_a} be the value of the outcome that would arise if the exposure is set to 1 and the version of the mediator is randomly selected from a distribution G_a . Likewise we will let Y_{0G_a} be the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from the distribution G_a .

Under assumptions (i*)–(iv*), we have the following results.

Result 1

Under assumptions (i*)–(iv*), the estimate, Q_1 of the natural indirect effect using data on only the mediator measurement M is equal to

$$Q_1 = E[Y_{1G_1} | c] - E[Y_{1G_0} | c].$$

Result 1 states that under the no-unmeasured-confounding assumptions (i*)–(iv*), the quantity we think we are estimating as the natural indirect effect, Q_1 using data only on our mediator measurement M , is in fact equal to a type of mediated effect. It is equal to the

effect of fixing the exposure to $A = 1$ and then randomly drawing a version of the mediator from the distribution G_1 versus the distribution G_0 , where G_1 first randomly selects a value of M from amongst those with $A = 1$, $C = c$ and G_0 first randomly selects a value of M from amongst those with $A = 0$, $C = c$ but then once the value of M has been selected each then randomly selects a version of the mediator from amongst those with $M = m$, $A = 1$, $C = c$. The expression is thus equal to a mediated effect, but with the versions of the mediator being drawn from the distribution of those with $M = m$, $A = 1$, $C = c$.

The next result gives an analogous result for the natural direct effect estimate but as will be seen the interpretation of the estimate in fact now involves two parts.

Result 2

Under assumptions (i*)–(iv*), the estimate, Q_2 , of the natural direct effect using data on only the mediator measurement M is equal to

$$Q_2 = (E[Y_{1G_0}|c] - E[Y_{0G_0}|c]) + (E[Y_{0G_0}|c] - E[Y_{0H_0}|c])$$

where Y_{0H_0} is the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from a distribution H_0 which first randomly selects a value of M from amongst those with $A = 0$, $C = c$ and then randomly selects a version of the mediator from amongst those with $M = m$, $A = 0$, $C = c$.

Contrary to Result 1 for the natural indirect effect, Result 2 for the natural direct effects says that if we proceed with the estimation of the natural direct effect using data only on the mediator measurement M (ignoring the different versions of the mediator) then what we think we are estimating as a natural direct effect is in fact the sum of two parts. The first part, $(E[Y_{1G_0}|c] - E[Y_{0G_0}|c])$ in the expression above, is in fact a direct effect. It compares fixing exposure to $A = 1$ with fixing exposure to $A = 0$ when setting the version of the mediator to a value randomly selected from the distribution G_0 . It is an effect of the exposure not through the mediator. The second part of the expression above for Q_2 , namely $(E[Y_{0G_0}|c] - E[Y_{0H_0}|c])$, is however not a direct effect but a type of mediated effect. It involves comparing what the outcome would be if exposure were fixed to 0 and the version of the mediator were randomly set to some value from the distribution G_0 versus one from distribution H_0 . Recall G_0 involves first randomly selecting a value of M from amongst those with $A = 0$, $C = c$ and then randomly selecting a version of the mediator from amongst those with $M = m$, $A = 1$, $C = c$; whereas H_0 involves first randomly selects a value of M from amongst those with $A = 0$, $C = c$ and then randomly selects a version of the mediator from amongst those with $M = m$, $A = 0$, $C = c$. The difference between these two distributions is in the second step wherein a version of the mediator is randomly selected from amongst those with $M = m$, $A = 1$, $C = c$ versus those with $M = m$, $A = 0$, $C = c$. Essentially then, what this second piece of the expression for Q_2 picks up is the effect of the exposure on outcome mediated through the portion of versions of the mediator that is not captured by mediator measurement M . The result is in some sense intuitive. If we ignore versions of the mediator in our mediation analysis, then the effect of the exposure on the outcome that is through versions of the mediator, K , but not through our mediator measure, M , will in fact be picked up by our direct effect estimate rather than our indirect effect estimate. This point should be taken into account in the estimation and interpretation of direct and indirect effects when multiple versions of the mediator exist but are not incorporated in the analysis. As described at greater length in the appendix, the quantities Q_1 and Q_2 do still add up to the total effect (as do the true natural direct and indirect effects) but

Q_1 and Q_2 are no longer equal to the true natural indirect and direct effects respectively since Q_2 captures part of the effect that is mediated.

An analogue somewhat similar to Result 2 in fact holds for controlled direct effects.

Result 3

Under assumptions (i*)–(iv*), the estimate, Q_3 , of the controlled direct effect using data on only the mediator measurement M is equal to

$$Q_3 = E[Y_{1F_1} - Y_{0F_1} | c] + E[Y_{0F_1} - Y_{0F_0} | c],$$

where F_1 is a random draw of a version from the distribution $P(k|A = 1, m, c)$ and F_0 is a random draw of a version from the distribution $P(k|A = 0, m, c)$.

The first part, $E[Y_{1F_1} - Y_{0F_1} | c]$ in the expression above, is a direct effect. It compares fixing exposure to $A = 1$ with fixing exposure to $A = 0$ when setting the version of the mediator to a value randomly selected from the distribution F_1 i.e. of those with $A = 1$, $M = m$, $C = c$. The second part of the expression, $E[Y_{0F_1} - Y_{0F_0} | c]$, is a type of mediated effect. It involves comparing what the outcome would be if exposure were fixed to 0 and the version of the mediator were randomly set to some value from the distribution F_1 versus one from distribution F_0 i.e. from amongst those with $A = 1$, $M = m$, $C = c$ versus those with $A = 0$, $M = m$, $C = c$.

Illustrations

Example 1. Dichotomization of the Mediator

When questions of mediation are of interest, investigators may, out of convenience, dichotomize a potential mediator. This raises questions with regard to what the interpretation of the resulting estimates of direct and indirect effects are under such dichotomization. Emsley et al.²⁴ for example, considered mediation in the Prevention of Suicide in Primary Care Elderly: Collaborative Trial (PROSPECT). They assessed whether the effect of randomized treatment (collaborative care management versus treatment as usual), A , on the score from the Hamilton Depression Scale, Y , was mediated by adherence to antidepressants. In their analyses they dichotomized adherence so that $M = 1$ indicated meeting with the relevant health specialist at least once and $M = 0$ indicated not. They assumed no interaction between the effect of A and M on Y and obtain an estimate of the direct effect of -3.15 (s.e. 0.82) and an estimate of the indirect effect of -0.49 . If there was indeed no exposure-mediator interaction, and if assumptions (i)–(iv) for the mediator M were satisfied, and adherence were truly captured by the binary indicator M , then their estimate of the direct effect would be equal to both the controlled and natural direct effect and their estimate of the indirect effect would be equal to the natural indirect effect. However, dichotomization of the adherence is likely a coarsening of the actual underlying adherence and will likely not entirely capture the effects of adherence on the depression score. Conceived of another way, there are multiple versions of the adherence indicator M , multiple values of adherence that correspond to e.g. $M = 1$.

Using the theory from the previous section we can consider the interpretation of the effect estimates of Emsley et al. when true adherence has been dichotomized for the purposes of the analysis. Suppose the no-unmeasured-confounding assumptions (i*)–(iv*) for the effects of treatment and adherence on depression scores hold; note that assumptions (i*) and (iii*) are guaranteed to hold by randomization of treatment. By Result 1, the indirect effect could

then be interpreted as the contrast of having the new treatment with the adherence indicator fixed to the value it would be with versus without the new treatment, in both cases choosing a particular value of actual adherence randomly from the distribution of those with that level of the adherence indicator, who actually received the new treatment, conditional on covariates (i.e. $A = 1$, $M = m$, $C = c$). By Result 2, the direct effect estimate is equal to the sum of (1) the contrast of the new treatment versus treatment as usual, in both cases with the adherence indicator set to the level it would have been with treatment as usual, with the specific actual adherence value randomly drawn from the distribution of those who actually received the new treatment, conditional on covariates, and (2) the effect of the new versus the standard treatment on the outcome mediated by adherence not captured by the adherence indicator itself.

Example 2. Unknown Versions of the Mediator

Caffo et al.²⁰ studied the extent to which the effect of cumulative lead dose, A , for organolead manufacturing workers on executive cognitive function test scores, Y , is mediated by white matter brain volumes, M , using data from 513 manufacturing workers in their study. Brain volume was measured using magnetic resonance imaging which captures only brain volume differences and not more subtle neurobiologic changes to brain structure. Caffo et al. control for a number of covariates, C , including age, education, smoking, and alcohol consumption, and use a regression based approach similar to that in Equation 1 above but assuming no $A \times M$ product term i.e. $\beta_3 = 0$ in regression models. They obtain an estimate of the direct effect of a 3.79 point decline (95% confidence interval [CI] = -7.40 to -0.18) in executive functioning cognitive test scores per 1- $\mu\text{g/g}$ increase in peak tibia lead exposure, controlling for white matter in brain regions associated with lead; the indirect effect of lead exposure as mediated through white matter brain volume was a 1.21 ($P = 0.01$) point decline in executive functioning cognitive test scores per 1- $\mu\text{g/g}$ increase in peak tibia lead exposure. If there was indeed no exposure-mediator interaction, and if assumptions (i)–(iv) for the mediator M were satisfied, and if there were no multiple versions of the mediator, then their estimate of the direct effect would be equal to both the controlled and natural direct effect and their estimate of the indirect effect would be equal to the natural indirect effect.

In this setting, however, there is no clear unambiguous way to conceive of hypothetical interventions on the mediator, brain volumes. It is not in fact clear at all what interventions would bring about changes to particular level. The versions of the mediator are effectively unknown. What then is the interpretation of the indirect effect estimate of Caffo et al.? Let us, for the time being, suppose the no-unmeasured-confounding assumptions (i*)–(iv*) for the exposure and the versions of the mediator with respect to the cognitive function outcome hold. Consider a one unit change in peak tibia lead exposure. By Result 1, the indirect effect would then be interpreted as the contrast of cognitive functioning test scores between two scenarios. In both scenarios, lead exposure is set to the higher level, but then in one scenario brain volume is fixed to the value it would have been with the higher lead exposure and in the other scenario brain volume is fixed to what it would have been under lower lead exposure, in both cases choosing a version of the mediator (to bring the specific brain volume about) randomly from the distribution of those who were actually at the higher level of lead exposure with that brain volume, conditional on covariates (i.e. $A = 1$, $M = m$, $C = c$). By Result 2, the direct effect estimate is equal to the sum of (1) the contrast of the higher versus lower level of lead exposure, in both cases with brain volumes set to the level it would have been under the lower exposure using a version of the mediator (to bring the specific brain volume about) randomly from the distribution of those who were actually at the higher level of lead exposure with that brain volume conditional on covariates and (2) the effect of the higher versus lower exposure on the outcome mediated by the various

versions of the manner in which the actual brain volume levels came about that is not captured by the brain volume measurement itself. The effects estimated by Caffo et al. may indeed have an interpretation as causal effects, albeit somewhat subtle and complicated, even though there are multiple ways a particular level of white matter brain volume could come about and even though we are not able to comprehensively characterize what these various versions might be. These interpretations, however, do assume that our no-unmeasured-confounding assumptions hold with respect to the underlying unknown versions of brain volumes, and with the versions unknown, this assumption will be difficult to assess in practice. We come back to this point in the discussion.

Discussion

In this paper we have considered the interpretation of estimates of direct and indirect effects when there are multiple versions of the mediator and these are not taken into account in the analysis. We have shown that what is estimated as a natural indirect effect does, under the relevant no-unmeasured-confounding assumptions, in fact capture a mediated effect with the version of the mediator randomly drawn from the distribution of versions that arises in the exposed population. However, we have also shown that what is estimated as a natural direct effect is in fact the sum of two quantities, one of which can be interpreted as a direct effect and the other of which is essentially the effect of the exposure on the outcome mediated by the effect of the exposure on version that is not captured by the mediator measurement. It is important to keep this latter point in mind in interpreting estimates of direct and indirect effects when multiple versions of the mediator may be present. Of course, if data were available on the version of the mediator that in fact led to each particular mediator measurement then analysis could proceed by simply replacing the mediator measurement by the version information and proceeding with the standard set of techniques from the causal inference literature^{1–11}. In most settings, however, this information will likely not be available, and the results here can then help inform the interpretation of the analysis.

The results here are important for two reasons. First, objections might be raised to the use of the counterfactual framework for mediation analysis in settings in which there are multiple versions of the mediator such that there is no single well-defined way to intervene on the mediator value of interest^{25,26}. The results here show that although this context complicates analysis and interpretation, the use of the counterfactual approach to mediation analysis is still relevant in this setting. The results here help address violations of the no-multiple-versions-of-treatment assumption; some parallels with potential violations of no-interference assumptions are given in the appendix. Second, however, in this context, what is estimated as a direct effect in fact picks up not only a direct effect but also a mediated effect through the versions of the mediator that is not captured by the mediator measurement. In many contexts, this may lead to an overestimate of direct effect and an underestimate of the indirect effect. Although this rule of thumb may be helpful in practice, it need not always hold. The effect mediated through the version of the mediator not captured by the mediator measurement need not necessarily be in the same direction as the natural indirect effect estimate through the mediator measurement. Indeed it has been shown elsewhere that when a mediator variable is dichotomized, without further assumptions, the natural indirect effect can be biased either upward or downward; and likewise with the natural direct effect.²⁷

Arguably the most substantial limitation of the approach described here is the no-unmeasured-confounding assumptions employed. This is an important limitation of analyses which seek to address questions of direct and indirect effects. Sensitivity analysis techniques have been developed to assess the extent to which violations in these assumptions would affect the qualitative and quantitative conclusions drawn from the estimates of the direct and indirect effects.^{7,9,11} However, in the context of multiple versions of the mediator, additional

complications arise insofar as the no-unmeasured-confounding assumptions being made are those with respect to the various versions of the mediator and not the mediator measurement itself. In many settings, an investigator may not be able to fully characterize all the possible versions of the mediator that may lead to a particular mediator measurement. In such cases, it will of course be very difficult to make assumption about no unmeasured confounding with any certainty when the variable with respect to which these assumptions are being made (the versions of the mediator) is not entirely known. Further work could perhaps be done in applying and extending sensitivity analysis techniques to this setting.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial Support: The research was supported by National Institutes of Health grant HD060696.

References

1. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3:143–155. [PubMed: 1576220]
2. Pearl, J. Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. San Francisco: Morgan Kaufmann; 2001. Direct and indirect effects; p. 411–420.
3. Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. Causal mediation analyses with rank preserving models. *Biometrics*. 2007; 63:926–934. [PubMed: 17825022]
4. van der Laan MJ, Petersen ML. Direct effect models. *International Journal of Biostatistics*. 2008:Article 23. [PubMed: 22462121]
5. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology*. 2009; 20:851–860. [PubMed: 19806060]
6. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface - Special Issue on Mental Health and Social Behavioral Science*. 2009; 2:457–468.
7. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology*. 2010; 172:1339–1348. [PubMed: 21036955]
8. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010; 21:540–551. [PubMed: 20479643]
9. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychological Methods*. 2010; 15:309–334. [PubMed: 20954780]
10. Ten Have TJ, Joffe M. A review of causal estimation of effects in mediation analyses. *Stat Methods Med Res*. 0962280210391076, first published on December 16, 2010. 10.1177/0962280210391076
11. Hafeman, D. *American Journal of Epidemiology*. Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. in press
12. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Evaluation Review*. 1981; 5:602–619.
13. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51:1173–1182. [PubMed: 3806354]
14. MacKinnon, DP. *An Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates; New York: 2008.
15. Rubin DB. Formal modes of statistical inference for causal effects. *J Statist Plan Inf*. 1990; 25(3): 279–292.

16. Robins JM, Greenland S. Comment on “Causal inference without counterfactuals. *Journal of the American Statistical Association*. 2000; 95:477–82.
17. Hernán MA. Invited commentary: Hypothetical interventions to define causal effects: afterthought or prerequisite? *American Journal of Epidemiology*. 2005; 162(7):618–620. [PubMed: 16120710]
18. Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*. 2009; 20:3–5. [PubMed: 19234395]
19. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009; 20(1):880–883. [PubMed: 19829187]
20. Caffo B, Chen S, Stewart W, Bolla K, Yousem D, Davatzikos C, Schwartz BS. Are brain volumes based on magnetic resonance imaging mediators of the associations of cumulative lead dose with cognitive function? *American Journal of Epidemiology*. 2008; 167:429–437.
21. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment; COBRA Preprint Series. p. Article 77 <http://biostats.bepress.com/cobra/ps/art77>
22. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011; 22:368–377. [PubMed: 21399502]
23. Valeri L, VanderWeele TJ. Technical Report. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros.
24. Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Stat Methods Med Res*. 2010; 19(3):237–270. [PubMed: 19608601]
25. MacLehose RF, Kaufman JS. *Epidemiology. The Wizard of Odds*. in press
26. VanderWeele TJ. Principal stratification: uses and limitations. *International Journal of Biostatistics*. 2011; 7:Article 28, 1–14. [PubMed: 21841939]
27. VanderWeele TJ, Hong G, Jones S, Brown J. Mediation and spillover effects in group-randomized trials: a case study of the 4R’s educational intervention. Revised for *Journal of the American Statistical Association*.
28. Rubin DB. Comment on: Neyman (1923) and Causal Inference in Experiments and. *Observational Studies Statistical Science*. 1990; 5:472–480.
29. Hong G, Raudenbush SW. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*. 2006; 101:901–910.
30. Sobel ME. What do randomized studies of housing mobility demonstrate? Causal Inference in the face of interference. *Journal of the American Statistical Association*. 2006; 101:1398–1407.
31. Rosenbaum PR. Interference between units in randomized experiments. *Journal of the American Statistical Association*. 2007; 102:191–200.
32. Hudgens MG, Halloran ME. Towards causal inference with interference. *Journal of the American Statistical Association*. 2008; 103:832–842. [PubMed: 19081744]
33. Tchetgen Tchetgen EJ, VanderWeele TJ. Estimation of causal effects in the presence of interference. *Statistical Methods in Medical Research - Special Issue on Causal Inference*. in press.
34. VanderWeele TJ, Tchetgen Tchetgen EJ. Bounding the infectiousness effect in vaccine trials. *Epidemiology*. 2011; 22:686–693. [PubMed: 21753730]
35. VanderWeele TJ, Tchetgen Tchetgen EJ. Effect partitioning under interference for two-stage randomized vaccine trials. *Statistics and Probability Letters*. 2011; 81:861–869. [PubMed: 21532912]

Appendix 1. Technical Details and Proofs

We let $X \perp\!\!\!\perp Y|Z$ denote that X is independent of Y conditional on Z . First consider the setting where there are not multiple versions of the mediator. Assumptions (i)–(iv) in the text, stated formally in terms of counterfactual independence, are: (i) $Y_{am} \perp\!\!\!\perp A|C$, (ii) $Y_{am} \perp\!\!\!\perp M\{A, C\}$, (iii) $M_a \perp\!\!\!\perp A|C$, (iv) $Y_{am} \perp\!\!\!\perp M_a^*|C$. Under these four assumptions the natural indirect and direct effects and controlled direct effect are identified respectively by^{2,6}:

$$\begin{aligned}
E[Y_{1M_1} - Y_{1M_0}] &= \sum_m E[Y|A=1, m, c] \{P(m|A=1, c) - P(m|A=0, c)\}, \\
E[Y_{1M_0} - Y_{0M_0}] &= \sum_m \{E[Y|A=1, m, c] - E[Y|A=0, m, c]\} P(m|A=0, c) \\
E[Y_{1M} - Y_{0M}] &= E[Y|A=1, m, c] - E[Y|A=0, m, c].
\end{aligned}$$

The right hand side of these three equalities were what was defined as Q_1 , Q_2 , Q_3 respectively in the text. These would be what would be estimated by techniques from mediation analysis using data only on A , M , Y and C . They are equal to the natural indirect and direct effects and the controlled direct effect respectively under assumptions (i)–(iv) above.

Now consider the setting in which there are multiple versions of the mediator. For each possible value m of M , there will be some set of possible interventions - some set of versions of the mediator - which would fix M to m ; we might denote this set by κ^m (and perhaps denote the various versions by 1, 2, 3, ..., etc.). For an individual with $M = m$ let K^m be the version that actually occurred so that M was at level m ; and if for that individual M was not level m then we might arbitrarily defined $K^m = 0$ because that there was no version of the mediator that set M to m since M was not level m . Following VanderWeele and Hernán²¹ for multiple versions of treatment for total effects, we let K denote a vector $(K^m : m \in \mathcal{M})$ where $\mathcal{M}$ is the support of M . The vector K will thus be all zeros except at one place, that corresponding to the value of the mediator that actually occurred for that individual. The variable M is effectively a coarsening of K and thus we will have that Y and A are independent of M conditional on K as in Figure 1. We can then let Y_{ak} be the counterfactual outcome that would have occurred had A been set to a and K to k and K_a as the value of K that would have occurred had A been set to a . The no-unmeasured-confounding assumptions for A and K with respect to Y can then be given as: (i*) $Y_{ak} \perp\!\!\!\perp A \mid C$, (ii*) $Y_{ak} \perp\!\!\!\perp K \mid \{A, C\}$, (iii*) $K_a \perp\!\!\!\perp A \mid C$, (iv*) $Y_{ak} \perp\!\!\!\perp K_a \mid C$. We can now prove the three results.

Proof of Results 1 and 2: Under assumptions (i*)–(iv*) we have that

$$\begin{aligned}
\sum_m E[Y|a, m, c] P(m|a^*, c) &= \sum_{k,m} E[Y|a, k, m, c] P(k|a, m, c) P(m|a^*, c) \\
&= \sum_k E[Y|A=a, k, c] \sum_m P(k|a, m, c) P(m|a^*, c) \\
&= \sum_k E[Y_{ak}|c] \sum_m P(k|a, m, c) P(m|a^*, c).
\end{aligned}$$

where the first equality follows by iterated expectations, the third by assumptions (i*) and (ii*). When $a = 1$, then, by the definition of G_{a^*} , this is also equal to:

$$\begin{aligned}
&= \sum_k E[Y_{1k}|c] \sum_m P(k|A=1, m, c) P(m|a^*, c) \\
&= \sum_k E[Y_{1k}|c] P(G_{a^*}=k|c) \\
&= E[Y_{1G_{a^*}}|c].
\end{aligned}$$

If data were only available on A , M , Y and C , but not on version, then the estimators used for the natural indirect effect would be consistent for

$$\begin{aligned}
Q_1 &= \sum_m E[Y|A=1, m, c] \{P(m|A=1, c) - P(m|A=0, c)\} \\
&= E[Y_{1G_1}|c] - E[Y_{1G_0}|c]
\end{aligned}$$

thus establishing Result 1. If data were only available on A , M , Y and C but not on version then the estimators used for the natural direct effect would be consistent for

$$\begin{aligned} Q_2 &= \sum_m \{E[Y|A=1, m, c] - E[Y|A=0, m, c]\} P(m|A=0, c) \\ &= \sum_m E[Y|A=1, m, c] P(m|A=0, c) - \sum_{m,k} E[Y|A=0, k, m, c] P(k|A=0, m, c) P(m|A=0, c) \\ &= E[Y_{1G_0}|c] - \sum_k E[Y_{0k}|c] \sum_m P(k|A=0, m, c) P(m|A=0, c) \\ &= E[Y_{1G_0}|c] - E[Y_{0H_0}|c] \\ &= (E[Y_{1G_0}|c] - E[Y_{0G_0}|c]) + (E[Y_{0G_0}|c] - E[Y_{0H_0}|c]) \end{aligned}$$

where the final lines follows by adding and subtracting $E[Y_{0G_0}|c]$, thus establishing Result 2. Note that assumption (iii*) is not strictly necessary for the proof of the result but is necessary for interpreting an expression such as $E[Y_{1G_1}|c] - E[Y_{1G_0}|c]$ as the effect of the exposure on the outcome mediated by version conditional on M and for interpreting $E[Y_{0G_0}|c] - E[Y_{0H_0}|c]$ as the effect of the exposure on the outcome mediated by version not captured by mediator measurement M . Moreover, although we did not make use of assumption (iv*), if there were an effect of the exposure that confounded the mediator-outcome relationship then assumption (ii*) would not hold conditional on a set of baseline covariate C and thus the derivations about would not follow.

Note that the estimators used for the natural indirect effect and the natural direct effect, Q_1 and Q_2 , do still add up to the total effect because

$$\begin{aligned} Q_1 + Q_2 &= \sum_m E[Y|A=1, m, c] \{P(m|A=1, c) - P(m|A=0, c)\} + \sum_m \{E[Y|A=1, m, c] - E[Y|A=0, m, c]\} P(m|A=0, c) \\ &= \sum_m E[Y|A=1, m, c] P(m|A=1, c) - \sum_m E[Y|A=0, m, c] P(m|A=0, c) \\ &= E[Y|A=1, c] - E[Y|A=0, c] \\ &= E[Y_1|c] - E[Y_0|c] \\ &= E[Y_1|c] - E[Y_0|c] \end{aligned}$$

where the third equality follows by iterated expectation, the fourth by consistency and the fifth by unconfoundedness of the exposure. Likewise the true natural direct and indirect effects add up to the total effect essentially by definition: $E[Y_{1M_1} - Y_{1M_0}|c] + E[Y_{1M_0} - Y_{0M_0}|c] = E[Y_{1M_1}|c] - E[Y_{0M_0}|c] = E[Y_1|c] - E[Y_0|c]$. However, the estimators used for the natural indirect effect and the natural direct effect, Q_1 and Q_2 , are no longer equal to the true natural indirect and direct effects respectively since Q_2 captures part of the effect that is mediated.

Proof of Result 3: If data were only available on A , M , Y and C but not on version then the estimators used for the controlled direct effect would be consistent for

$$\begin{aligned} Q_3 &= E[Y|A=1, m, c] - E[Y|A=0, m, c] \\ &= \sum_k E[Y|A=1, k, m, c] P(k|A=1, m, c) - \sum_k E[Y|A=0, k, m, c] P(k|A=0, m, c) \\ &= \sum_k E[Y|A=1, k, c] P(k|A=1, m, c) - \sum_k E[Y|A=0, k, c] P(k|A=0, m, c) \\ &= \sum_k E[Y_{1k}|c] P(k|A=1, m, c) - \sum_k E[Y_{0k}|c] P(k|A=0, m, c) \\ &= \sum_k E[Y_{1k} - Y_{0k}|c] P(k|A=1, m, c) + \sum_k E[Y_{0k}|c] \{P(k|A=1, m, c) - P(k|A=0, m, c)\} \end{aligned}$$

where the second to last equality follows by assumptions (i*) and (ii*) and the final equality follows by adding and subtracting $\sum_k E[Y_{0k}|c] P(k|A=1, m, c)$. From the definition of F_a we then also have that this is equal to:

$$\begin{aligned}
&= \sum_k E[Y_{1k} - Y_{0k} | c] P(F_1 = k | c) + \sum_k E[Y_{0k} | c] P(F_1 = k | c) - \sum_k E[Y_{0k} | c] P(F_0 = k | c) \\
&= E[Y_{1F_1} - Y_{0F_1} | c] + E[Y_{0F_1} - Y_{0F_0} | c]
\end{aligned}$$

thus establishing Result 3.

Another way to view the controlled direct effect estimand is thus as two parts where the first part is the controlled direct effect for the exposure on the outcome not through version, standardized by the distribution of the versions of the mediator amongst those with $A = 1$, $M = m$, $C = c$. The second part is a comparison of the counterfactual $E[Y_{0k} | c]$ standardized by the distribution of versions amongst those with $A = 1$, $M = m$, $C = c$ versus amongst those with $A = 0$, $M = m$, $C = c$. This is once again picks up an effect of the exposure on the version of the mediator not through the mediator measure M .

Appendix 2. Parallels Between Violations of the No-Multiple-Versions-of-Treatment Assumption and the No-Interference Assumption

In the standard counterfactual or potential outcomes framework, two assumptions are generally made so as to have well defined counterfactuals: no multiple versions of treatment and no interference. The first of these has been the focus of this paper. These two assumptions are sometimes referred to as stability assumptions. They are the two parts of what Rubin, in describing this potential outcomes framework called “SUTVA”, the Stable Unit Treatment Value Assumption.²⁸ The first assumption, no multiple versions of treatment, is that in whatever way the treatment is administered it results in the same outcome for a particular individual. The second assumption, no interference, is that the exposure or treatment of one individual does not affect the outcomes of another. These two assumptions were traditionally made in all of the causal inference literature so that the potential outcomes notation was well-defined and justified. In many settings these assumptions, however, are likely unreasonable, and work has recently been undertaken to address violations of both the no-multiple-versions-of-treatment assumption^{19,21,22} and the no-interference assumption^{29–35}. In discussing the estimation of total effect in the presence of multiple versions of treatment, Van-derWeele and Hernán²¹, noted several common features between the no-multiple-versions-of-treatment and no-interference assumptions. First, both the no-interference assumption and the no-multiple-versions-of-treatment assumption are concealed by the standard potential outcomes notation and the assumptions are generally made implicitly by the use of such notation. Second, with suitable modification, violations of both assumptions can be accommodated by expanding notation. Third, in certain settings, violations of assumptions can in fact be ignored, at least if only testing for causal effects is of interest. And fourth, once notation has been introduced to expand the potential outcomes framework in order to accommodate violations of either assumption, this can give rise to new effect of interest. See VanderWeele and Hernan²¹ for further discussion of these points.

Here we have further extended causal inference under “multiple versions” to settings in which mediation - direct and indirect effects - are of interest and where there are multiple versions of the mediator. In other work, mediation analysis techniques have been extended to settings that allow for interference and spillover effects for the effects of a mediator.²⁷ A comparison with this work gives rise to a new parallel between multiple versions and interference. In the context of interference, it was shown that, under certain assumptions, if there is interference such that the value of the mediator for one individual may affect the outcome of another, and this interference is ignored, then what is estimated as a natural direct effect in fact captures both the true natural direct effect and also a spillover mediator

effect i.e. some of the mediated effect is picked up by the natural direct effect estimate.²⁷ Likewise, here, we have shown that if there are multiple versions of the mediator, and these are ignored in the analysis then what is estimated as a natural direct effect in fact captures a direct effect and also a mediated effect through version that is not captured by the actual mediator measurement; once again some of the mediated effect is picked up by the natural direct effect estimate. The two results are in some sense analogous. As theory for causal inference under violations of the no-multiple-versions or no-interference assumptions expands, additional parallels and correspondences may emerge. In contexts such as that considered here, these results, can facilitate the interpretation of the results of analyses in complicated and more realistic settings.

Figure 1.
Diagram indicating relations between exposure A , underlying versions of the mediator K , the mediator measure M , the outcome Y and the confounding variables C .