Web Material for "When Is the Difference Method Conservative for Assessing Mediation?"

Web Appendix 1: Mediation Analysis Based on Counterfactual Framework

One of the advantages of the counterfactual approach to mediation analysis is that it allows for the decomposition of a total effect into a direct effect and an indirect effect even when nonlinearities and interactions are present (1, 2).

We will let Y_a and M_a denote respectively, the value of outcome and mediator that would have been observed if we set $A = a$, and we let Y_{am} be the value of outcome that would have been observed if we set $A = a$ and $M = m$. Therefore $Y_{aM_{a'}}$ is the value of outcome that would have been observed if we set $A = a$ and $M = M_{a'}$. We make the consistency assumption which is commonly used in causal inference literature (3, 4), i.e. $Y_{am} = Y$ when $A = a$, $M = a$ and $Y_a = Y, M_a = M$ when $A = a$.

We will follow the exposition of Pearl (2) to propose the definition of the total effect, the natural direct and indirect effects. On the difference scale, the average total effect, conditional on $X = x$, comparing exposure levels a with a' $(a > a')$, is defined by $RD^{TE}_{a,a'|x} = E(Y_a - Y_{a'} \mid x)$, which compares the average outcome in subgroup $X = x$ if A had been set to a with the average outcome in subgroup $X = x$ if A had been set to a'. On the odds ratio scale, the average total effect, conditional on $X = x$, comparing exposure levels a with a', is defined by

$$OR^{TE}_{a,a'|x} = \frac{P(Y_a = 1 \mid x)/\{1 - P(Y_a = 1 \mid x)\}}{P(Y_{a'} = 1 \mid x)/\{1 - P(Y_{a'} = 1 \mid x)\}},$$

which compares the odds of outcome $Y = 1$ in subgroup $X = x$ if A had been set to a

with the odds of outcome $Y = 1$ in subgroup $X = x$ if A had been set to a.

There are two classes of direct effects: "controlled" effects and "natural" effects. The average controlled direct effect conditional on x on the risk difference scale is defined as $E(Y_{am} - Y_{a'm} \mid x)$, which captures the effect of the exposure with an intervention on the mediator to fix it to a specific value. We can also define the natural direct and indirect effects on either the difference or the odds ratio scale. On the difference scale, the natural direct effect, conditional on x, comparing the effect of the exposure levels a and a' while fixing the mediator at the level it would have naturally been under some reference condition for the exposure, $A = a'$, is defined by $RD^{NDE}_{a,a'|x} = E(Y_{aM_{a'}} - Y_{a'M_{a'}} \mid x)$. The natural direct effect on the risk difference scale compares the average outcome in subgroup $X = x$ if A had been set to a and M had been set to $M_{a'}$ with the average outcome in subgroup $X = x$ if A had been set to a' and M had been set to $M_{a'}$. Thus the natural direct effect captures the effect of exposure on the outcome via pathways that do not involve the mediator M.

The natural indirect effect, conditional on x, comparing the effect of the mediator at levels M_a and $M_{a'}$ while fixing the exposure at level a, is defined by $RD^{NIE}_{a,a'|x} = E(Y_{aM_a} - Y_{aM_{a'}} \mid x)$. The natural indirect effect on the risk difference scale compares the average outcome in subgroup $X = x$ if A had been set to a and M had been set to M_a with the average outcome in subgroup $X = x$ if A had been set to a and M had been set to $M_{a'}$.

On the odds ratio scale, the natural direct effect is defined by

$$OR^{NDE}_{a,a'|x} = \frac{P(Y_{aM_{a'}} = 1 \mid x)/\{1 - P(Y_{aM_{a'}} = 1 \mid x)\}}{P(Y_{a'M_{a'}} = 1 \mid x)/\{1 - P(Y_{a'M_{a'}} = 1 \mid x)\}},$$

which compares the odds of $Y = 1$ in subgroup $X = x$ if A had been set to a and M had been set to $M_{a'}$ with the odds of $Y = 1$ in subgroup $X = x$ if A had been set to a' and M had been set to $M_{a'}$. And the natural indirect effect is defined by

$$OR^{NIE}_{a,a'|x} = \frac{P(Y_{aM_a} = 1 \mid x)/\{1 - P(Y_{aM_a} = 1 \mid x)\}}{P(Y_{aM_{a'}} = 1 \mid x)/\{1 - P(Y_{aM_{a'}} = 1 \mid x)\}},$$

which compares the odds of $Y = 1$ in subgroup $X = x$ if A had been set to a and M had been set to M_a with the odds of $Y = 1$ in subgroup $X = x$ if A had been set to a and M had been set to $M_{a'}$. Therefore, the natural indirect effect captures the effect of exposure on the outcome through the mediator M.

On the difference scale, the natural direct and indirect effects have the property that the total effect decomposes into a natural direct and indirect effect:

$$\begin{aligned} RD^{TE}_{a,a'|x} &= E(Y_a - Y_{a'} \mid x) = E(Y_{aM_a} - Y_{aM_{a'}} \mid x) + E(Y_{aM_{a'}} - Y_{a'M_{a'}} \mid x) \\ &= RD^{NDE}_{a,a'|x} + RD^{NIE}_{a,a'|x}. \end{aligned}$$

A similar property holds on the odds ratio scale, that the odds ratio for the total effect decomposes into a product of a natural direct effect and indirect effect:

$$\begin{aligned} OR^{TE}_{a,a'|x} &= \frac{P(Y_a = 1 \mid x)/\{1 - P(Y_a = 1 \mid x)\}}{P(Y_{a'} = 1 \mid x)/\{1 - P(Y_{a'} = 1 \mid x)\}} \\ &= \frac{P(Y_{aM_{a'}} = 1 \mid x)/\{1 - P(Y_{aM_{a'}} = 1 \mid x)\}}{P(Y_{a'M_{a'}} = 1 \mid x)/\{1 - P(Y_{a'M_{a'}} = 1 \mid x)\}} \cdot \frac{P(Y_{aM_a} = 1 \mid x)/\{1 - P(Y_{aM_a} = 1 \mid x)\}}{P(Y_{aM_{a'}} = 1 \mid x)/\{1 - P(Y_{aM_{a'}} = 1 \mid x)\}} \\ &= OR^{NDE}_{a,a'|x} \cdot OR^{NIE}_{a,a'|x}, \end{aligned}$$

which is equivalent to

$$\log(OR^{TE}_{a,a'|x}) = \log(OR^{NDE}_{a,a'|x}) + \log(OR^{NIE}_{a,a'|x}).$$

Under certain assumptions, the total effect, the natural direct and indirect effects can be identified with observed data. We will follow the exposition of VanderWeele

(5) and VanderWeele and Vansteelandt (6) on the identification assumptions proposed by Pearl (2). These assumptions were presented to identify natural direct and indirect effects on the risk difference scale, and also presented to be sufficient on the odds ratio scale. Imai, Keele and Yamamoto (7) propose alternative assumptions to identify natural direct and indirect effects. But on the causal diagrams with non-parameteric structural equation models (3), the two sets of assumptions are equivalent (8). We will let $A \perp\!\!\!\perp B \mid C$ to denote A is independent of B conditional on C.

To identify the total effect, the general assumption is that conditional on some set of measured covariates X, the effect of exposure A on outcome Y is unconfounded, i.e.

$$Y_a \perp\!\!\!\perp A \mid X. \tag{1}$$

In practice, one should collect data on a sufficiently rich set of covariates X to control for confounding. Under this assumption, $RD^{TE}_{a,a'|x}$ and $OR^{TE}_{a,a'|x}$ can be estimated from the data as follows:

$$RD^{TE}_{a,a'|x} = E(Y \mid a, x) - E(Y \mid a', x),$$
$$OR^{TE}_{a,a'|x} = \frac{P(Y=1 \mid a,x)/\{1-P(Y=1 \mid a,x)\}}{P(Y=1 \mid a',x)/\{1-P(Y=1 \mid a',x)\}}.$$

Identifying the natural direct and indirect effects entails stronger assumptions and we will present and illustrate them in two parts.

First, we need the assumptions that conditioning on the set of covariates X suffices to control for confounding of both the exposure-outcome and the mediator-outcome relations, i.e.

$$Y_{am} \perp\!\!\!\perp A \mid X, \tag{2}$$

$$Y_{am} \perp\!\!\!\perp M \mid \{A, X\}. \tag{3}$$

Assumption 2 is similar to the assumption needed for identifying total effects, and Assumption 3 requires that, conditional on $\{A, X\}$, there is no unmeasured confounding for the mediator-outcome relation. Assumptions 2 and 3 together mean that the set of covariates X suffice to control for the confounding of exposure-outcome relation, and also to control for the confounding of mediator-outcome relation. These two assumptions are restrictive in the sense that there can not be any additional covariates other than X to confound the mediator-outcome relation.

However, Assumptions 2 and 3 are not sufficient to identify the natural direct and indirect effects, and we need the following two additional assumptions:

$$M_a \perp\!\!\!\perp A \mid X, \tag{4}$$

$$Y_{am} \perp\!\!\!\perp M_{a'} \mid X. \tag{5}$$

Assumption 4 means that conditional on X, there is no unmeasured confounding of the exposure-mediator relation. Assumption 5 can be interpreted as that there is no mediator-outcome confounder affected by the exposure.

Assumptions 2 to 5 are not necessarily satisfied even in randomized experiments, therefore, in practice we need to rely on scientific knowledge to judge these assumptions and sensitivity analysis should be conducted when some assumptions are violated (5, 7).

Under Assumptions 2 to 5, the natural direct and indirect effects on the difference scale and the odds ratio scale can be estimated as:

$$
\begin{aligned}
RD^{NDE}_{a,a'|x} &= \int P(Y=1 \mid a,m,x)P(m \mid a',x)dm - \int P(Y=1 \mid a',m,x)P(m \mid a',x)dm,\\
OR^{NDE}_{a,a'|x} &= \frac{\int P(Y=1 \mid a,m,x)P(m \mid a',x)dm/\{1-P(Y=1 \mid a,m,x)P(m \mid a',x)dm\}}{\int P(Y=1 \mid a',m,x)P(m \mid a',x)dm/\{1-P(Y=1 \mid a',m,x)P(m \mid a',x)dm\}},\\
RD^{NIE}_{a,a'|x} &= \int P(Y=1 \mid a,m,x)P(m \mid a,x)dm - \int P(Y=1 \mid a,m,x)P(m \mid a',x)dm,\\
OR^{NIE}_{a,a'|x} &= \frac{\int P(Y=1 \mid a,m,x)P(m \mid a,x)dm/\{1-P(Y=1 \mid a,m,x)P(m \mid a,x)dm\}}{\int P(Y=1 \mid a,m,x)P(m \mid a',x)dm/\{1-P(Y=1 \mid a,m,x)P(m \mid a',x)dm\}}.
\end{aligned}
$$

Since Assumption 1 is implied by Assumption 2 to 5, the confounding assumptions (a)-(d) in the main text is referred to as Assumptions 2 to 5.

Web Appendix 2: Proof of the Results

Proof of Result 1.

$$\begin{aligned} RD^{TE}_{a,a'|x} &= E(Y \mid a,x) - E(Y \mid a',x) \\ &= (\phi_0 + \phi_1 a + \phi_2^\top x) - (\phi_0 + \phi_1 a' + \phi_2^\top x) = \phi_1(a-a'). \end{aligned}$$

Under Assumptions 2-5, according to model 2 the natural direct effect on the risk difference scale is given by:

$$\begin{aligned} RD^{NDE}_{a,a'|x} &= \int P(Y=1 \mid a,m,x)P(m \mid a',x)dm - \int P(Y=1 \mid a',m,x)P(m \mid a',x)dm \\ &= \{\theta_0 + \theta_1 a + \theta_2 E(M \mid a',x) + \theta_3^\top x\} - \{\theta_0 + \theta_1 a' + \theta_2 E(M \mid a',x) + \theta_3^\top x\} \\ &= \theta_1(a-a'). \end{aligned}$$

From the property that the total effect decomposes into a natural direct and indirect effect, we then have $RD^{NIE}_{a,a'|x} = (\phi_1 - \theta_1)(a-a')$. Result 1 then follows if we let $a=1$ and $a'=0$.

We use the following Lemma to prove Result 2:

Lemma 1. *Let f and g be functions with n real-valued arguments such that both f and g are non-decreasing in each of their arguments. If $X = (X_1, \cdots, X_n)$ is a multivariate random variable with n components such that each component is independent of the other components, then $Cov\{f(X), g(X)\} \geq 0$*

Proof of Lemma 1. See Theorem 2.1 of Esary, Proschan and Walkup (9).

Proof of Result 2. According to model 3, if Assumption 1 holds, we can get

$$\begin{aligned} \log(OR^{TE}_{a,a'|x}) &= \log\left\{\frac{P(Y=1 \mid a,x)/\{1-P(Y=1 \mid a,x)\}}{P(Y=1 \mid a',x)/\{1-P(Y=1 \mid a',x)\}}\right\} \\ &= (\phi_0 + \phi_1 a + \phi_2^\top x) - (\phi_0 + \phi_1 a' + \phi_2^\top x) = \phi_1(a-a'). \end{aligned}$$

Thus we have $OR^{TE}_{a,a'|x} = e^{\phi_1(a-a')}$. Under Assumption 2 to 5, from model 4, we can estimate $OR^{NDE}_{a,a'|x}$ as:

$$OR^{NDE}_{a,a'|x} = (\frac{1}{A} - 1)/(\frac{1}{B} - 1),$$

where

$$\begin{aligned} A &= \int P(Y = 0 \mid a, m, x)P(m \mid a', x)dm \\ &= \int \frac{1}{1 + e^{\theta_0+\theta_1 a+\theta_2 m+\theta_3^\top x}} \cdot P(M = m \mid a', x)dm \\ &= E_{M|a',x}\left(\frac{1}{1 + e^{\theta_0+\theta_1 a+\theta_2 M+\theta_3^\top x}}\right), \end{aligned}$$

and

$$\begin{aligned} B &= \int P(Y = 0 \mid a', m, x)P(m|a', c)dm \\ &= \int \frac{1}{1 + e^{\theta_0+\theta_1 a'+\theta_2 m+\theta_3^\top x}} \cdot P(M = m \mid a', x)dm \\ &= E_{M|a',x}\left(\frac{1}{1 + e^{\theta_0+\theta_1 a'+\theta_2 M+\theta_3^\top x}}\right). \end{aligned}$$

Then, we have

$$\begin{aligned} & e^{\theta_1(a-a')}(\frac{1}{B} - 1) - (\frac{1}{A} - 1) \\ = & \frac{A \cdot e^{\theta_1(a-a')} - B}{AB} - e^{\theta_1(a-a')} \\ = & \frac{E_{M|a',x}\left\{\frac{e^{\theta_1(a-a')}}{1+e^{\theta_0+\theta_1 a+\theta_2 M+\theta_3^\top x}} - \frac{1}{1+e^{\theta_0+\theta_1 a'+\theta_2 M+\theta_3^\top x}}\right\}}{AB} - (e^{\theta_1(a-a')} - 1) \\ = & (e^{\theta_1(a-a')} - 1)(\frac{C}{AB} - 1), \end{aligned}$$

where $C = E_{M|a',x}\left\{\frac{1}{(1+e^{\theta_0+\theta_1 a+\theta_2 M+\theta_3^\top x})(1+e^{\theta_0+\theta_1 a'+\theta_2 M+\theta_3^\top x})}\right\}$.

Since $\frac{1}{1+e^{\theta_0+\theta_1 a+\theta_2 m+\theta_3^\top x}}$ and $\frac{1}{1+e^{\theta_0+\theta_1 a'+\theta_2 m+\theta_3^\top c}}$ are both non-increasing or non-decreasing in m, according to Lemma 1, $C - AB = \mathrm{Cov}_{M|a',x}\left(\frac{1}{1+e^{\theta_0+\theta_1 a+\theta_2 M+\theta_3^\top x}}, \frac{1}{1+e^{\theta_0+\theta_1 a'+\theta_2 M+\theta_3^\top x}}\right) \geq 0$.

If $\theta_1 \geq 0$, we have $(e^{\theta_1(a-a')} - 1)(\frac{C}{AB} - 1) \geq 0$, and then $OR^{NDE}_{a,a'|c} \leq e^{\theta_1(a-a')}$. From the property that on the odds ratio scale, the total effect decomposes into a product of a natural direct effect and indirect effect , we have $OR^{NIE}_{a,a'|x} = e^{\phi_1(a-a')}/OR^{NDE}_{a,a'|c} \geq e^{(\phi_1-\theta_1)(a-a')}$. Result 2 then follows if we let $a = 1$ and $a' = 0$.

Web Appendix 3: Result for the confidence interval of the natural indirect effect based on the difference method

Result 3. *(a) If $\theta_1 \geq 0$, and the lower 95% confidence bound of the difference method is positive, then the lower 95% confidence bound of the natural indirect effect is positive. (b) if $P(\theta_1 > 0, \phi_1 - \theta_1 > 0) \geq 0.95$, then the lower 95% confidence bound of the natural indirect effect is positive.*

Proof of Result 4. (a) From Result 2, we have

$$P(\log(OR^{NIE}_{a,a'|c}) > 0) \geq P(\phi_1 - \theta_1 > 0) > 0.95.$$

(b) From Result 2, we have

$$\begin{aligned} & P(\log(OR^{NIE}_{a,a'|c}) > 0) = P(\log(OR^{NIE}_{a,a'|c}) > 0, \theta_1 \geq 0) \\ \geq \;\; & P(\phi_1 - \theta_1 > 0, \theta_1 \geq 0) \geq 0.95. \end{aligned}$$

Web Appendix 4: Other Qualitative Conclusions About the Natural Indirect Effect

First, we need M to come from an exponential family which is defined as follows:

Definition 1. *We say that Y is from an exponential family if its probability density function has the form*

$$f(y;\theta,\phi) = e^{\frac{y\theta+b(\theta)}{a(\phi)}+c(y,\phi)}$$

Then, we have

Result 4. *Under Assumptions 2 to 5, suppose M comes from the exponential family with $E(M \mid a,x) \geq E(M \mid a',x)$, then*

(a) if model 2 holds with $\theta_2 \geq 0$, $RD^{NIE}_{a,a'|x} \geq 0$;

(b) if models 4 holds with $\theta_2 \geq 0$, $log(OR^{NIE}_{a,a'|x}) \geq 0$.

The following Lemmas are useful for our derivation:

Lemma 2. *Suppose X_1, X_2 are from the exponential family with the same ϕ, then*

$$\theta_{x_1} \geq \theta_{x_2} \Leftrightarrow E(X_1) \geq E(X_2) \Leftrightarrow P(X_1 > x) \geq P(X_2 > x).$$

Proof of Lemma 2. See Theorem 3.4.1 (ii) of Lehmann and Romano (10).

Lemma 3. *If $E(Y|a,m,x)$ is non-decreasing(non-increasing) in m, and $P(M > m|a,x)$ is non-decreasing(non-increasing) in a for all m, then $\int E(Y \mid a,m,x)\{P(M = m \mid a,x) - P(M = m \mid a',x)\}dm \geq 0$.*

Proof of Lemma 3. See Lemma 1 of VanderWeele and Robins (11).

Proof of Result 3. If model 2 holds, we have

$$RD^{NIE}_{a,a'|x} = \int E(Y \mid a,m,x)\{P(M=m \mid a,x) - P(M=m \mid a',x)\}dm.$$

Since $\theta_2 \geq 0$ implies $E(Y|a,m,x)$ is non-decreasing in m, and according to Lemma 2 we can get $P(M > m|a,x)$ is non-decreasing(non-increasing) in a for all m from $E(M \mid a,x) \geq E(M \mid a',x)$. From Lemma 3, we have $RD^{NIE}_{a,a'|x} \geq 0$.

If model 4 holds, we have

$$OR^{NIE}_{a,a'|x} = \frac{\int P(Y=1 \mid a,m,x)P(m \mid a,x)dm/\{1-\int P(Y=1 \mid a,m,x)P(m \mid a,x)dm\}}{\int P(Y=1 \mid a,m,x)P(m \mid a,x)dm/\{1-\int P(Y=1 \mid a,m,x)P(m \mid a',x)dm\}}.$$

Since $\theta_2 \geq 0$ implies $P(Y=1|a,m,x)$ is non-decreasing in m, and according to Lemma 2 we can get $P(M > m|a,x)$ is non-decreasing(non-increasing) in a for all m from $E(M \mid a,x) \geq E(M \mid a',x)$. From Lemma 3, we have $\int P(Y=1 \mid a,m,x)P(m \mid a,x)dm \geq \int P(Y=1 \mid a,m,x)P(m \mid a',x)dm$, thus $log(OR^{NIE}_{a,a'|x}) \geq 0$.

The following Corollary follows immediately from Theorem 4.

Corollary 1. *Under Assumptions 2 to 5, if model 4 and a linear model* $E(M|a,x) = \beta_0 + \beta_1 a + \beta_2' x$ *hold with* $\theta_2 \geq 0$ *and* $\beta_1 \geq 0$*, then* $log(OR^{NIE}_{a,a'|x}) \geq 0$.

References

[1] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology.* 1992;3(2):143–155.

[2] Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Inteligence.* San Francisco, CA: Morgan Kaufmann, 411–420.

[3] Pearl J. *Causality: Models, Reasoning, and Inference.* Cambridge, United Kingdom: Cambridge University Press, 2nd ed., 2009.

[4] VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology.* 2009;20(6):880–883.

[5] VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Eur J Epidemiol.* 2009;20(1):18–26.

[6] VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface.* 2009;2:457–468.

[7] Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci.* 2010;25(1):51–71.

[8] Shpitser I, VanderWeele TJ. A complete graphical criterion for the adjustment formula in mediation analysis. *Int J Biostat.* 2011;7(1):1–24.

[9] Esary JD, Proschan F, Walkup DW. Association of random variables, with applications. *Ann Math Sci.* 1967;38(5):1466–1474.

[10] Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. Springer, 2006.

[11] VanderWeele TJJ, Robins JMM. Properties of monotonic effects on directed acyclic graphs. *The Journal of Machine Learning Research* 2009;10(699-718).