Causal organic direct and indirect effects: closer to Baron and Kenny

Judith J. Lok
Department of Mathematics and Statistics, Boston University
March 13, 2019

Abstract

Baron and Kenny (1986, 80,433 Google Scholar citations) proposed estimators of direct and indirect effects: the part of a treatment effect that is mediated by a covariate and the part that is not. Subsequent work on natural direct and indirect effects provides a formal causal interpretation. Natural direct and indirect effects use cross-worlds counterfactuals: outcomes under treatment with the mediator “set” to its value without treatment. Organic direct and indirect effects (Lok 2016) avoid cross-worlds counterfactuals, using “organic” interventions on the mediator while keeping the initial treatment fixed at “treatment”. Organic direct and indirect effects apply also to settings where the mediator cannot be “set”. In linear models where there is no treatment-mediator interaction, both organic and natural indirect effects lead to the same estimators as in Baron and Kenny (1986). In this article, I propose organic interventions on the mediator that keep the initial treatment fixed at “no treatment”, leading to an alternative version of organic direct and indirect effects. I show that the product method, proposed in Baron and Kenny (1986), holds for these new direct and indirect effects even if there is treatment-mediator interaction. Furthermore, I argue that this alternative organic indirect effect is more relevant for drug development than the traditional natural or organic indirect effect.

1 Introduction: direct and indirect effects

Direct and indirect effects decompose the effect of a treatment $A$ on outcome $Y$ into:

- a part that is mediated through covariate $M$ (the indirect effect)
- a part that is not (the direct effect).

For example, treatment $A$ could be a blood pressure lowering medication, and the outcome $Y$ could be whether a person had a heart attack. The mediation question could then be: How much of the effect of the blood pressure lowering
medication is mediated by its effect on blood pressure, and how much (if any) by other pathways? Another mediation question is (Lok 2016): How much of the effect of the effect of antidepressants on depressive symptoms is mediated through participants’ expectations? Another example is in HIV research. HIV infected mothers can infect their infants, either at birth or through breast feeding. This HIV transmission can partly be prevented. The mediation question could then be (Sperling et al. 1996): How much of the effect of AZT treatment on mother-to-child transmission is mediated through the effect of AZT on HIV-1 RNA?

The seminal article on the topic of mediation analysis, Baron and Kenny (1986), has over 80,000 citations in Google Scholar, many of them from the last 10 years. Mediation analysis is especially important in the health sciences, like epidemiology and psychology. Knowing the type of assumptions under which these analyses are valid is paramount. Such assumptions are not yet well-established. From Robins and Richardson (2011): “The nature of the relationship between the sentence expressing these causal conclusions and the statistical computer calculations performed on the strings of numbers has been obscure.”

One of the disadvantages of many current causal interpretations of mediation analysis (see e.g. Pearl 2001, Imai et al. 2010, VanderWeele 2009, and Tchetgen-Tchetgen 2011) lies in their reliance on quantities such as: the outcome under treatment but with the mediator “set” to its value without treatment. These are called cross-world counterfactuals, since they rely on two simultaneous but different interventions, “treatment” and “no treatment”, which never occur concurrently. Then, obviously, the identifying assumptions are also cross-world assumptions, and these have often been disputed (Robins and Richardson 2011), and have led some researchers to entirely forego causal mediation analysis.

Lok (2016) has proposed an approach to causal mediation analysis that does not rely on cross-worlds quantities or cross-world assumptions. This approach, organic direct and indirect effects, is outlined in Section 3. This article proposes to adapt organic direct and indirect effects further, and proposes direct and indirect effects that are

- Closer to the original approach to mediation analysis by Baron and Kenny (1986), and
- More useful in identifying the indirect effect of treatments, to select the most promising treatments for further investigation in randomized trials.

This article thus has three purposes: to bring organic direct and indirect effects to the attention of epidemiologists, to bring causal mediation analysis closer to the original approach in Baron and Kenny (1986), and to make causal mediation analysis more useful in the selection of treatments for randomized trials.
2 Setting and notation

We start with the situation where treatment $A$ is randomized. Denote the pre-treatment common causes of the mediator $M$ and the outcome $Y$ by $C$. As is usual in the mediation literature, we start by assuming that there are no post-treatment common causes of the mediator $M$ and the outcome $Y$. This can be relaxed, see [Lok (2017)]. Throughout, the subscript 0 indicates “without treatment” and the subscript 1 indicates “under treatment”. The DAG in Figure 1 illustrates the mediation set-up. Notice the absence of an arrow from $C$ into $A$ since treatment is randomized.

![Figure 1: DAG summarizing the data](image)

We adopt the usual Consistency Assumption relating the observed data to the counterfactual data:

**Assumption:** (Consistency). On $A = 1$, $M = M_1$ and $Y = Y_1$. On $A = 0$, $M = M_0$ and $Y = Y_0$.

3 Definition of causal direct and indirect effects

3.1 Natural direct and indirect effects

Historically (e.g. [Pearl (2001), Imai et al. (2010), VanderWeele (2009), Robins and Richardson (2011) and Tchetgen-Tchetgen (2011)]), causal direct and indirect effects have often been defined in terms of the following counterfactuals: the outcomes under treatment had the mediator been “set” to the mediator without treatment: $Y_{1,M_0}$. There are two issues with these counterfactual outcomes $Y_{1,M_0}$:

- Under treatment, $M_0$ is not observed. Thus, even if we could set the mediator, how to set it to $M_0$ under treatment ($A = 1$)?

- How to set the mediator is usually left unanswered, so the outcomes $Y_{1,M_0}$ are undefined in many practical situations. [Cole and Frangakis (2009)] provide an appealing example: “There are many competing ways to assign (hypothetically) a body mass index of 25 kg/m² to an individual, and each of them may have a different causal effect on the outcome”.


Based on the outcomes $Y_{1,M_0}$, most current causal approaches to mediation analysis focus on natural direct and indirect effects:

- **Natural direct effect**: $E(Y_{1,M_0} - Y_0)$ (not mediated: $M \equiv M_0$).

- **Natural indirect effect**: $E(Y_1 - Y_{1,M_0})$ (mediated through $M$: $M_1$ vs $M_0$).

To estimate the natural direct and indirect effect, since treatment $A$ is randomized, estimation of $EY_1$ and $EY_0$ is standard. One can focus on estimating $E(Y_{1,M_0})$. Under (strong!) conditions, with $C$ all pre-treatment common causes of the mediator $M$ and the outcome $Y$: the “Mediation Formula” holds (see e.g. Pearl (2001)):

$$E(Y_{1,M_0}) = \int_{(m,c)} E[Y|M = m, C = c, A = 1] f_{M|C=c,A=0}(m) f_C(c) dm dc.$$  

This result is obviously appealing: first, the pre-treatment $C$ come about, then $M$ with its distribution under $A = 0$, and then $Y$ with its distribution under $A = 1$. The conditions under which this was proven were strong, and relied on cross-worlds assumptions. In addition, most causal mediation approaches need many counterfactual outcomes: not only $Y_{1,M_0}$, also all $Y_{a,m}$: the outcomes with the treatment set to $a$ and the mediator set to $m$. When the mediator is a second medication, such as aspirin in Pearl (2001), these outcomes $Y_{a,m}$ are conceivable. However, if the mediator is a patient characteristic or covariate, as is often the case in applications, these outcomes $Y_{a,m}$ are often not conceivable.

Under additional conditions (linear models and no exposure-mediator interaction), the resulting estimators are the same as in Baron and Kenny (1986). Under these conditions, the causal inference literature thus adds a causal interpretation to the estimators from Baron and Kenny (1986).

### 3.2 Organic direct and indirect effects: an intervention-based approach

Lok (2016) has shown that the mediation formula holds under much weaker assumptions, for quantities that are not cross-world quantities, and without cross-world assumptions, as follows.

In the following, $I$ will be an intervention on the mediator that does not affect pre-treatment common causes $C$ of the mediator $M$ and the outcome $Y$. $M_{1,I=1}$ and $Y_{1,I=1}$ will denote the mediator and the outcome under treatment and under intervention $I$ on the mediator.

**Definition: (Organic intervention).** An intervention $I$ is an organic intervention with respect to $C$ if

$$M_{1,I=1} \mid C = c \sim M_0 \mid C = c$$  

(1)
and

\[ Y_{1,I=1} \mid M_{1,I=1} = m, C = c \sim Y_1 \mid M_1 = m, C = c, \]  

(2)

where \( \sim \) indicates having the same distribution.

Equation (1) says that \( I \) “holds the mediator at its distribution under no treatment”: given \( C \), there is no difference in the distribution of the mediator under treatment and intervention \( I \) and the distribution of the mediator under no treatment. Equation (2) says that \( I \) “has no direct effect on the outcome”: under treatment, whether mediator = \( m \) came about due to \( I \) or without \( I \) does not affect a person’s prognosis. Or, “how the mediator came about is irrelevant”. That is,

\[ P(Y_{1,I=1} \leq y \mid M_{1,I=1} = m, C = c) = P(Y_1 \leq y \mid M_1 = m, C = c) \]

for all \( y \). \[ \text{Lok (2016)} \] showed that for equation (2) to make sense, one needs all pre-treatment common causes of the mediator \( M \) and the outcome \( Y \) in \( C \).

Without \( C \), the statement “mediator under treatment equals \( m \)” likely implies a different prognosis under intervention \( I \) (\( M_{1,I=1} \)) versus without intervention \( I \) (\( M_1 \)), because \( M_1 \) and \( M_{1,I=1} \) may be related in a different way to \( C \), which we assumed predicts the outcome.

An example of an “organic” intervention could be the following. \( A = 1 \) could be a blood pressure lowering medicine, \( M \) a person’s blood pressure, and \( Y \) the occurrence of a heart attack. The mediation question: does \( A = 1 \) have a direct effect on heart attacks? It could be, e.g., that \( A = 1 \) lowers blood pressure by 10, on average, without changing the shape of the blood pressure distribution. \( I \) should then be an intervention, in the treated, that increases the blood pressure by 10, on average, without changing the shape of the blood pressure distribution. Then, \( M_{1,I=1} \sim M_0 \). \( I \) could be salt in a dosage dependent on \( C = c \), perhaps a dosage with some specific distribution given \( C = c \). The effect of salt on heart attacks is believed to be through its effect on blood pressure (see for example the CDC website, http://www.cdc.gov/vitalsigns/Sodium/index.html); thus, one can hope that

\[ Y_{1,I=1} \mid M_{1,I=1} = m, C = c \sim Y_1 \mid M_1 = m, C = c \]

for this intervention with salt, \( I \).

Whether an intervention \( I \) is organic can (and should) be discussed with subject matter experts. In the following, we will draw conclusions about the effect of organic interventions \( I \).

**Definition:** ("Organic" direct and indirect effects \[ \text{Lok (2016)} \]):

- \( EY_1 - EY_{1,I=1} \): organic indirect effect of treatment \( A \) based on \( I \). ((Treatment= 1 for both \( Y_1 \) and \( Y_{1,I=1} \), so mediated)).

- \( EY_{1,I=1} - EY_0 \): organic direct effect of treatment \( A \) based on \( I \). ((Mediator same distribution for \( Y_{1,I=1} \) and \( Y_0 \), so not mediated)).
Lok (2016) has shown that the organic direct and indirect effects do not depend on the intervention $I$, as long as $I$ is organic. Lok (2016) has also shown that if $C$ has all pre-treatment common causes of the mediator $M$ and the outcome $Y$, the choice of pre-treatment common causes $C$ does not affect the organic direct and indirect effects; see also Web-appendix D.

Lok (2016) has proven that under the usual conditions to identify natural direct and indirect effects, organic direct and indirect effects generalize natural direct and indirect effects. Provided that $M_{1,I=1} = M_0$ exists, the usual “cross-worlds” assumption implies that the intervention $I$ that sets the mediator to $M_0$, $M_{1,I=1} = M_0$, is organic! This approach also provides a proof of the mediation formula for natural direct and indirect effects under conditions that are somewhat weaker than usual, see Web-appendix D. Lok (2016) has also proven that the organic direct and indirect effects generalize the direct and indirect effects proposed by Didelez et al. (2006), which are based on randomizing and then setting the mediator.

As for natural direct and indirect effects, under randomized treatment, estimation of $E[Y_0]$ and $E[Y_1]$ is standard. For $E[Y_{1,I=1}]$, Lok (2016) provides the mediation formula (see also Section 4) under only the above conditions.

The mediation formula thus provides the same identification result as previously found for the natural direct and indirect effects studied by previous authors, of note Pearl (2001), VanderWeele (2009), Imai et al. (2010), and Tchetgen-Tchetgen (2011). The resulting expression is in terms of observable quantities only. All ingredients can be estimated using standard methods. The contribution of Lok (2016) to this literature is, that the definition and therefore the interpretation of direct and indirect effects as well as the conditions under which estimators for these effects are valid can be considerably relaxed.

This intervention based approach answers questions about the effect of interventions, and what one might expect from interventions that satisfy certain conditions.

4 Organic direct and indirect effects: towards an alternative definition

Section 3.2 combined organic interventions on the mediator with “treatment”, that is, with $A = 1$. That approach follows analogies with natural direct and indirect effects. However, the product method from Baron and Kenny (1986) does not work for these causal estimands in linear models when there is treatment–mediator interaction in the outcome model; it only works when there is no treatment–mediator interaction in the outcome model.

In this article, we propose to combine organic interventions on the mediator with “no treatment”, that is, with $A = 0$. This proposal has important advantages:

1. The product method from Baron and Kenny (1986) does work for linear models regardless of whether there is treatment–mediator interaction in
the outcome model.

2. Organic indirect effects can be estimated with:
   (a) The distribution of the mediator under treatment and under “no treatment”, and
   (b) The relation between the mediator and the outcome under “no treatment”.

   This has important advantages for selecting new treatments with promising indirect effects for clinical trials (!)

In the following, let $M_{0,I=1}$ and $Y_{0,I=1}$ be the mediator and the outcome under no treatment and under intervention $I$ on the mediator. We redefine an organic intervention on the mediator as follows:

**Definition: (Organic intervention).** An intervention $I$ is an organic intervention with respect to $C$ if

\[ M_{0,I=1} \mid C = c \sim M_1 \mid C = c \quad (3) \]

and

\[ Y_{0,I=1} \mid M_{0,I=1} = m, C = c \sim Y_0 \mid M_0 = m, C = c. \quad (4) \]

Equation (3) says that $I$ “changes the distribution of the mediator to that under treatment”: given $C$, there is no difference in the distribution of the mediator under intervention $I$ and the distribution of the mediator under treatment. Equation (4) says that $I$ “has no direct effect on the outcome”: whether mediator = $m$ came about due to $I$ or without $I$ does not affect a person’s prognosis. Or, “how the mediator came about is irrelevant”.

**Definition: (New “organic” direct and indirect effects):**

- $EY_{0,I=1} - EY_0$: **organic indirect effect** of treatment $A$ based on $I$. ((Treatment = 0 for both $Y_0$ and $Y_{0,I=1}$, so mediated)).
- $EY_1 - EY_{0,I=1}$: **organic direct effect** of treatment $A$ based on $I$. ((Mediator same distribution for $Y_1$ and $Y_{0,I=1}$, so not mediated)).

Combining organic interventions with “no treatment” provides useful information on what to expect from a treatment that:

1. Affects the mediator the same way as the treatment does.
2. Has no direct effect on the outcome.

This is arguably more relevant to drug development than the previously defined causal indirect effects; for an HIV transmission example see Lok (2016) and Web-appendix C.
Similar to Lok (2016), see Web-appendix A it follows that the Mediation Formula holds for organic interventions:

**Theorem (Organic direct and indirect effects: the Mediation Formula for randomized data).** Under randomized treatment, consistency, and the definition of organic interventions, the following holds for an intervention $I$ that is organic with respect to $C$:

$$E(Y_{0,I=1}) = \int_{(m,c)} E[Y|M = m, C = c, A = 0] f_{M|C=c,A=1}(m)f_C(c)dm dc.$$

Note: the mediation formula does not depend on the choice of organic intervention $I$, confirming that the type of organic intervention does not affect the magnitude of the organic direct and indirect effect.

Under linear models and in the absence of treatment-mediator interaction in the outcome model, the new definition of organic direct and indirect effects leads to the same results as in Lok (2016). It can also be shown (see Web-appendix B) that with the new definition, the product method from Baron and Kenny (1986) works for linear models regardless of whether there is treatment-mediator interaction in the outcome model. This is in contrast with both Lok (2016) and previous work on causal mediation analysis, both for natural direct and indirect effects (Pearl (2001), Imai et al. (2010), VanderWeele (2009), Robins and Richardson (2011) and Tchetgen-Tchetgen (2011)) as for the approach advocated in Didelez et al. (2006).

### 4.1 Selecting new treatments with promising indirect effects for clinical trials

From the Mediation Formula, the new organic indirect effect $EY_{0,I=1} - EY_0$ equals

$$\int_{(m,c)} E[Y|M = m, C = c, A = 0] (f_{M|C=c,A=1}(m) - f_{M|C=c,A=0}(m)) f_C(c)dm dc.$$

Thus, the organic indirect effect can be estimated with:

(a) The distribution of the mediator under treatment, $A = 1$, and under “no treatment”, $A = 0$.

(b) The expectation of the outcome given the mediator and pre-treatment covariates $C$ under “no treatment”, $A = 0$ (only under $A = 0$).

Expanding mediation analysis in this way is of particular interest because it allows estimation of the indirect effect of potential new treatments that are being developed to affect a mediator (e.g., Frank and Hargreaves (2003)) without measuring the outcomes under the new treatments, because the distribution of the
outcome given the mediator under treatment is not needed. This is important for treatment development, where there are often many candidate treatments for which the effects on biomarkers have been conjectured or established. It is however important to note that evaluating this indirect effect is just a first step in treatment evaluation, because it only estimates the indirect effect. For this indirect effect to be the total effect, equation (4) would need to hold for the candidate treatment $A$ taking the role of $I$, which can only be confirmed by measuring the outcomes under the candidate treatment. Thus, the candidate treatments need to be evaluated in a randomized trial, but randomized trials could be reserved for treatments with the most promising indirect effect.

5 Application: selecting HIV cure treatments with promising indirect effects for clinical trials

Now that ART has rendered HIV a chronic disease, a substantial body of HIV research is focusing on HIV eradication, or cure, aimed at long-term ART-free HIV remission. For the development of medications for HIV eradication, estimating the indirect effect of potential new treatments is particularly important. Trials are being developed to interrupt the current standard of care, ART, in HIV infected patients, to investigate the effect of on-ART biomarkers on the time to viral rebound: the time at which the HIV viral load in the blood of the patients is above a pre-specified level (Etemad et al. (2015), Li et al. (2015)). This will lead to estimates of the dependence of the outcome given $M$ and $C$ under no treatment, needed to evaluate the indirect effect of equation (5).

ART interruption trials have to be carried out with extreme care, because ART interruption carries significant risks (Strategies for Management of Antiretroviral Therapy (SMART) Study Group et al. (2006), Li et al. (2015), Li et al. (2016)). Because there are many potential new drugs, it is advantageous to carry out ART interruption trials for only the most promising drugs (Ghosn and Delaugerre (2015)). If a potential new drug is designed to affect a biomarker $M$ which has an effect on the time to viral rebound $Y$, equation (5) provides an estimate of the indirect effect of the potential new drug mediated by $M$, without interrupting ART.

Through our collaboration with Dr. Bosch (Center for Biostatistics in AIDS Research, Harvard TH Chan School of Public Health) and Dr. Li (Brigham and Women’s hospital, Boston), we will have access to the data from completed ACTG trials described in Li et al. (2016): data on > 100 HIV infected patients with ART interruptions, and the following biomarker measurements: cell associated HIV RNA and single-copy plasma HIV RNA. For the purpose of this illustrative analysis, we categorize these as above versus below the assay lower limit of detection, and also for illustrative purposes, we ignore potential post-treatment common causes of the mediator and the outcome. We illustrate our methods by estimating, from on-ART data, the indirect effect of potential
treatments that affect these biomarkers on the time to viral rebound after ART interruption.

6 Discussion

Both for natural direct and indirect effects and for the direct and indirect effects proposed in [Didelez et al. (2006)], one needs to be able to set mediator to any specific value. For “organic” direct and indirect effects, one only needs to be able to affect the distribution of the mediator, and that is often what potential new treatments will aim to accomplish. Organic direct and indirect effects as introduced in [Lok (2016)] generalize natural direct and indirect effects.

To answer clinical questions, it is often useful to combine interventions on the mediator with “no treatment”, rather than with “treatment”. Our new definition of organic direct and indirect effects, which combines interventions on the mediator with “no treatment”, no longer generalizes natural direct and indirect effects; it is however closer to the original definition proposed by [Baron and Kenny (1986)]. The newly defined organic indirect effects can be estimated with information on

1. The distribution of the mediator under treatment, \( A = 1 \), and under “no treatment”, \( A = 0 \).
2. The expectation of the outcome given the mediator and pre-treatment covariates \( C \) under “no treatment”, \( A = 0 \) (only under \( A = 0! \)).

This is useful for selecting potential new treatments with a promising indirect effect for further evaluation in clinical trials.

Analogous to [Lok (2016)], the newly proposed organic direct and indirect effects can also be estimated from observational data, provided that all confounders have been measured; see Web-appendix. [Incorporating post-treatment common causes of the mediator and the outcome is an interesting topic for future research; some preliminary results can be found in Lok (2017).]

Acknowledgements

This research was supported by the National Science Foundation DMS 1810837. The content is solely the responsibility of the author and does not necessarily represent the official views of the NSF.

References

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

Cole, S. R. and C. E. Frangakis (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 20(1), 3–5.
DeGruttola, V. G., P. Clax, D. L. DeMets, G. J. Downing, S. S. Ellenberg, L. Friedman, M. H. Gail, R. Prentice, J. Wittes, and S. L. Zeger (2001). Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop. Controlled Clinical Trials 22, 485–502.

Didelaez, V., A. P. Dawid, and S. Geneletti (2006). Direct and indirect effects of sequential treatments. In Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, pp. 138–146. Arlington, VA: AUAI Press: arxiv preprint arXiv 1206.6840.

Etemad, B., H. Ahmed, E. Aga, R. Bosch, J. W. Mellors, D. Kuritzkes, M. Para, R. T. Gandhi, and J. Li (2015). The size of the active HIV reservoir predicts timing of viral rebound. In Conference on Retroviruses and Opportunistic Infections (CROI), Seattle, WA.

Frank, R. and R. Hargreaves (2003). Clinical biomarkers in drug discovery and development. Nature Reviews Drug Discovery 2(7), 566–580.

Ghosn, J. and C. Delaugerre (2015). Can we avoid treatment interruption studies in the search for an HIV cure? AIDS 29, 1575–1577.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science 25, 51–71.

Li, J. Z., B. Etemad, H. Ahmed, E. Aga, R. J. Bosch, J. W. Mellors, D. R. Kuritzkes, M. M. Lederman, M. Para, and R. T. Gandhi (2016). The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. AIDS 30(3), 343–353.

Li, J. Z., D. M. Smith, and J. W. Mellors (2015). The need for treatment interruption studies and biomarker identification in the search for an HIV cure. AIDS 29, 1429–1432.

Lok, J. J. (2016). Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. Statistics in Medicine 35(22), 4008–4020.

Lok, J. J. (2017). Organic direct and indirect effects with post-treatment common causes of mediator and outcome. Technical report. arXiv: http://arxiv.org/abs/1510.02753.

Pearl, J. (2000). Causality. Models, reasoning, and inference. Cambridge: Cambridge University Press.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the 17th annual conference on uncertainty in artificial intelligence (UAI-01), pp. 411–442. San Francisco: Morgan Kaufmann.
Robins, J. M. and T. S. Richardson (2011). Alternative Graphical Causal Models and the Identification of Direct Effects. In P. Shrout, K. Keyes, and K. Ornstein (Eds.), *Causality and psychopathology: finding the determinants of disorders and their cures*, Chapter 6, pp. 1–52. Oxford University Press.

Sperling, R. S., D. E. Shapiro, R. W. Coombs, J. A. Todd, S. A. Herman, G. D. McSherry, M. J. O’Sullivan, R. B. VanDyke, E. Jiminez, C. Rouzioux, P. M. Flynn, and J. L. Sullivan (1996). Maternal viral load, Zidovudine treatment, and the risk of transmission of Human Immunodeficiency Virus type 1 from mother to infant. *New England Journal of Medicine* 335(22), 1621–1629.

Strategies for Management of Antiretroviral Therapy (SMART) Study Group, W. M. El-Sadr, J. D. Lundgren, J. D. Neaton, F. Gordin, D. Abrams, R. C. Arduino, and et al. (2006). CD4+ count-guided interruption of antiretroviral treatment. *New England Journal of Medicine* 355, 2283–2296.

Tchetgen-Tchetgen, E. J. (2011). On causal mediation analysis with a survival outcome. *The International Journal of Biostatistics* 7(1), 1–38.

VanderWeele, T. J. (2009). Marginal Structural Models for the estimation of direct and indirect effects. *Epidemiology* 20(1), 18–26.

**A The Mediation Formula for organic direct and indirect effects**

In this Web-appendix, we prove the Mediation Formula for the newly proposed organic direct and indirect effects.

**Theorem (Organic direct and indirect effects: the Mediation Formula for randomized data).** Under randomized treatment, consistency, and the definition of organic interventions, the following holds for an intervention $I$ that is organic with respect to $C$: 

$$E(Y_{0,I=1}) = \int_{(m,c)} E[Y|M = m, C = c, A = 0] f_{M|C=c,A=1}(m)f_C(c) dm dc.$$
Proof:

\[
E(Y_{0,t=1}) = E(E[Y_{0,t=1}|M_{0,t=1}, C])
\]

\[
= \int_{(m,c)} E[Y_{0}|M_{0} = m, C = c] f_{M_0,C=c}(m) dm f_C(c) dc
\]

\[
= \int_{(m,c)} E[Y_{0}|M_{0} = m, C = c] f_{M_1,C=c}(m) dm f_C(c) dc
\]

\[
= \int_{(m,c)} E[Y|M = m, C = c, A = 0] f_{M_1,C=c,A=1}(m) dm f_C(c) dc
\]

\[
= \int_{(m,c)} E[Y|M = m, C = c, A = 0] f_{M|C=c,A=1}(m) f_C(c) dm dc.
\]

The second equation follows because of the definition of conditional expectations. The third equation follows because of the two parts of the definition of organic interventions. The fourth equation follows because under randomized treatment,

\[A \perp \perp (Y_0, M_0) | C\] and \[A \perp M_1 | C\].

The last equation follows from the Consistency Assumption.

B New definition of organic direct and indirect effects: the product method of Baron and Kenny (1986): proof

In this Web-appendix, we show that the product method of Baron and Kenny (1986) holds under linear models, regardless of whether there is an interaction between treatment and the mediator in the outcome model.

Assumptions for product method.

\[M = \alpha_0 + \alpha_1 C + \alpha_2 A + \epsilon_M,\]

with \(\epsilon_M \perp \perp A|C\), and

\[Y = \beta_0 + \beta_1 C + \beta_2 A + \beta_3 M + \beta_4 AM + \epsilon_Y,\]

with \(E[\epsilon_Y|M, A, C] = 0\).

Theorem (product method for new organic indirect effect). Under the above assumptions for the product method, the newly proposed organic indirect effect is equal to

\[E(Y_{0,t=1}) - EY_0 = \beta_3 \alpha_2.\]
Proof:

\[ E(Y_{0,I=1}) - EY_0 \]
\[ = \int_{(m,c)} E(Y|M = m, A = 0, C = c) \]
\[ = \int_{(m,c)} \left( f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m) \right) f_C(c) dm dc \]
\[ = \beta_3 \int_{(m,c)} m \left( f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m) \right) f_C(c) dm dc \]
\[ + \int_c (\beta_0 + \beta_1 c) \int_m \left( f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m) \right) dm dc \]
\[ = \beta_3 \int_{(m,c)} m f_{M|A=0,C=c}(m - \alpha_2) f_C(c) dm dc \]
\[ - \beta_3 \int_{(m,c)} m f_{M|A=0,C=c}(m) f_C(c) dm dc + 0 \]
\[ = \beta_3 \int_{(\tilde{m},c)} (\tilde{m} + \alpha_2) f_{M|A=0,C=c}(\tilde{m}) f_C(c) d\tilde{m} dc \]
\[ - \beta_3 \int_{(m,c)} m f_{M|A=0,C=c}(m) f_C(c) dm dc \]
\[ = \beta_3 \alpha_2. \]

For the first equality we used the Mediation Formula. For the fourth equality we used that \( f_{M|A=1,C=c}(m) = f_{M|A=0,C=c}(m - \alpha_2) \). For the fifth equality we substituted \( \tilde{m} = m - \alpha_2 \). That finishes the proof.

C Usefulness of combining “no treatment” with intervention on the mediator: mother-to-child transmission of HIV/AIDS

HIV infected mothers can transmit the HIV virus to their infants. The effect of AZT treatment on mother-to-child transmission of HIV-1 is surprisingly large, given the limited effect of AZT mono-therapy on HIV-1 RNA (DeGruttola et al. [2001]). Less than 20% of the effect of AZT on mother to child transmission can be explained through the effect of AZT on HIV-1 RNA (Sperling et al. [1996]).

What is the likely effect on mother-to-child transmission of a potential new treatment that has the same effect on HIV-1 RNA as AZT but no “direct” effect on mother to child transmission? In this case, the outcome \( Y \) is an indicator “newborn baby is HIV infected”. The mediator \( M \) is HIV-1 RNA. \( I \) is an intervention that, without AZT, causes the distribution of HIV-1 RNA, \( M_{0,I=1}, \)
to be the same as under AZT; the potential new treatment. The quantity of interest is then $EY_{0,I=1} - EY_0$. Note: this quantity is different from the usual indirect effect, since it combines an intervention $I$ on the mediator with “no treatment”, instead of with “treatment”.

The natural direct and indirect effect don’t have a lot of meaning in this case, since the HIV viral load in the mother’s blood cannot be “set”. Once we will be able to set the HIV viral load, we will set it to 0. Thus, for a treatment like AZT, organic direct and indirect effects are more natural than their natural counterparts.

Combining organic interventions with “no treatment” provides useful information on what to expect from a treatment that:

1. Affects the HIV viral load in the mother’s blood the same way as AZT does.
2. Has no direct effect on mother-to-child transmission.

## D  Uniqueness of organic direct and indirect effects

**Definition:** (common cause). $X$ is not a common cause of mediator and outcome given $C$ if either equation (6) or equation (7) holds:

$$X \perp \!\!\!\!\!\!\perp M_0 \mid C \quad \text{and} \quad X \perp \!\!\!\!\!\!\perp M_1 \mid C$$  \hspace{1cm} (6)

or

$$X \perp \!\!\!\!\!\!\perp Y_0 \mid M_0, C.$$  \hspace{1cm} (7)

In graphical language: $X$ is a common cause of the mediator and the outcome if in a DAG that has $C, X, M,$ and $Y$, there is an arrow from $X$ to $M$, and there is a direct arrow from $X$ to $Y$. This definition of common cause is in line with, for example, [Pearl (2000)](#).

Now, let $I^C$ be an intervention that is organic with respect to $C$ and let $I^{\tilde{C}}$ be an intervention that is organic with respect to $\tilde{C}$. Assume that $C$ is not a common cause of the mediator and the outcome given $\tilde{C}$, and that $\tilde{C}$ is not a common cause of the mediator and the outcome given $C$; hence there are 4 different cases, with either (6) or (7) holding for $C$ and $\tilde{C}$, respectively. Similar to Lok (2016), it can be shown that under any of those 4 different cases,

$$E \left( Y_{0,I^{\tilde{C}}=1} \right)$$

$$= \int_{(m,\tilde{c},c)} E \left[ Y_0 \mid M_0 = m, \tilde{C} = \tilde{c}, C = c \right] f_{M_1,C=\tilde{c},C=c}(m) f_{\tilde{C},C}(\tilde{c},c) dm \, d\tilde{c} \, dc.$$

Because of symmetry, it follows that also

$$E \left( Y_{0,I^C=1} \right)$$

$$= \int_{(m,\tilde{c},c)} E \left[ Y_0 \mid M_0 = m, \tilde{C} = \tilde{c}, C = c \right] f_{M_1,C=\tilde{c},C=c}(m) f_{\tilde{C},C}(\tilde{c},c) dm \, d\tilde{c} \, dc.$$
But then, $E(Y_{0,Ic=1}) = E(Y_{0,Ic=1})$. That finishes the proof.

E  A weaker identifiability condition for natural direct and indirect effects

Consider the situation where the mediator under treatment can be set to its value without treatment, and there is consensus about the closest possible world to this one where this can be accomplished. In this case, the natural direct and indirect effects are well-defined. Now let $M_{1,I=1} = M_0$; this is an intervention $I$ on the mediator. If all $Y_{a,m}$ “exist”, this intervention $I$ is an organic intervention if:

$$Y_{1,I=1|M_{1,I=1} = m, C = c} \sim Y_1|M_1 = m, C = c,$$

or equivalently, since in this particular example $M_{1,I=1} = M_0$ and all $Y_{a,m}$ exist,

$$Y_{1,m|M_0 = m, C = c} \sim Y_{1,m}|M_1 = m, C = c.$$  \hspace{1cm} (8)

Analogous to Lok (2016), this follows e.g. under the usual conditions for identification of natural direct and indirect effects. Then, the natural direct and indirect effects are the same as the organic direct and indirect effects. Still, equation (8) is a cross-worlds assumption, with $Y_{1,m}$ conditional on $M_0$. This is no surprise, since the definition of natural direct and indirect effects relies on cross-worlds quantities. Still, the above relaxes the usual assumptions for identification of natural direct and indirect effects.

F  Observational data

With observational data, the definition of organic direct and indirect effects should not change. Here we show that an identification result holds for observational data similar to the identification result for randomized data, provided that $C$ has all common causes of the mediator and the outcome.

There may exist baseline covariates $Z$ (beyond the common causes $C$ of the mediator and the outcome) that need to be included in the analysis in order to eliminate confounding:

**Assumption:** (No Unmeasured Confounding).

$$A \perp (Y_0, M_0) \mid C, Z \quad \text{and} \quad A \perp Y_1 \mid C, Z \quad \text{and} \quad A \perp M_1 \mid C, Z.$$  

**Theorem:** (Organic direct and indirect effects: the Mediation Formula for observational data). Assume No Unmeasured Confounding, Consistency, intervention $I$ is organic with respect to $C$, and given $C, Z$ is not a common cause
of the mediator and the outcome (see Web-appendix D). Then

$$E(Y_{0, I=1}) = \int_{(m, c, z)} E[Y|M = m, C = c, Z = z, A = 0] f_M|C=c, Z=z, A=1(m) f_{C,Z}(c, z) dm \, d(c, z).$$

The proof of this Mediation Formula is similar to the proofs in Web-appendix A and in Lok (2016).