Causal Organic Indirect and Direct Effects: Closer to Baron and Kenny, and Related to Surrogate Outcomes

Judith J. Lok
Associate Professor of Mathematics and Statistics
Boston University, jjlok@bu.edu

First International Workshop on Interactive Causal Learning, 6/11/2022

Joint work with Ronald J. Bosch
Center for Biostatistics in AIDS Research
Harvard T.H. Chan School of Public Health

Supported by: NSF DMS 1854934, NIH/NIAID UM1 AI068634 and AI068636 (but not the view of NSF or NIH)

Motivation: indirect and direct effects

Indirect and direct effects decompose the effect of a treatment A on outcome Y into:

- part that is mediated through covariate M (*indirect effect*)
- part that is not (*direct effect*).

Example:
A blood pressure lowering medication,
outcome Y: heart attack.

How much of the effect of the blood pressure lowering medication A on heart attacks Y is mediated by the effect of A on blood pressure M, and how much (if any) by other pathways?

Motivation: indirect and direct effects

Indirect and direct effects decompose the effect of a treatment A on outcome Y into:

- part that is mediated through covariate M (*indirect effect*)
- part that is not (*direct effect*).

Example: there are several potential/pre-clinical HIV curative treatments A, typically targeting the HIV reservoir M, which may prolong the time to viral rebound after ART interruption, Y.

What is the effect of an HIV curative treatment A on viral rebound Y mediated through the HIV reservoir M?

Motivation for mediation analysis

The seminal article on mediation analysis, Baron and Kenny (1986), has 111,810 citations in Google Scholar, many from the last 10 years.

Mediation analysis is very important in the health sciences, like epidemiology and psychology.

Knowing (and relaxing) the type of assumptions under which these analyses are valid/causal is important.

Setting and notation

First: randomized treatment A.

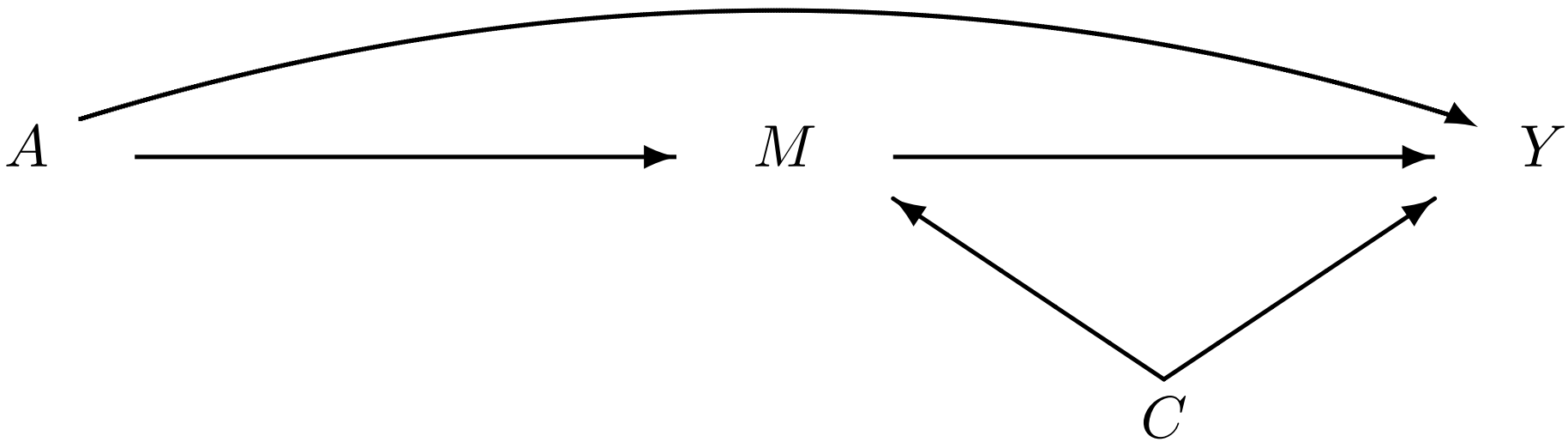
Pre-treatment common causes of mediator M and outcome Y: C.

No post-treatment common causes of mediator M and outcome Y. Can be relaxed; not this talk.

Superscript $^{(0)}$ indicates "without treatment".

Superscript $^{(1)}$ indicates "under treatment".

Figure 1: Causal diagram summarizing the data



Motivation: definition of causal indirect and direct effects

Causal indirect and direct effects are often defined in terms of the following counterfactuals:

The outcomes under treatment had the mediator been "set" to the mediator without treatment: $Y^{(1,M^{(0)})}$.

Two issues with these counterfactual outcomes $Y^{(1,M^{(0)})}$:

1. How to set the mediator is usually left unanswered $\Rightarrow$ outcomes undefined in many practical situations.

Cole and Frangakis (2009) provide an illustrative example: "There are many competing ways to assign (hypothetically) a body mass index of 25 kg/m^2 to an individual, and each of them may have a different causal effect on the outcome".

Motivation: definition of causal indirect and direct effects

Causal indirect and direct effects are often defined in terms of the following counterfactuals:

The outcomes under treatment had the mediator been "set" to the mediator without treatment: $Y^{(1,M^{(0)})}$.

Two issues with these counterfactual outcomes $Y^{(1,M^{(0)})}$:

2. Under treatment, $M^{(0)}$ not observed. $\Rightarrow$ even if we could set the mediator, how to set it to $M^{(0)}$ under treatment $(A = 1)$?

Motivation: definition of causal indirect and direct effects

Natural indirect effect: $E\left(Y^{(1)} - Y^{(1,M^{(0)})}\right)$ (*mediated* through M).

Natural direct effect: $E\left(Y^{(1,M^{(0)})} - Y^{(0)}\right)$ (*not mediated:* $M \equiv M^{(0)}$).

See e.g. Robins and Greenland (1992); Pearl (2001); Imai et al. (2010); VanderWeele (2009, 2015); Robins and Richardson (2011); Tchetgen-Tchetgen (2011).

Motivation: identification result for natural indirect and direct effects

Suppose randomized treatment A. Then estimation of $EY^{(1)}$ and $EY^{(0)}$ standard. Left to estimate: $E\left(Y^{(1,M^{(0)})}\right)$. Under (strong!) conditions, with C all pre-treatment common causes of mediator M and outcome Y: "**Mediation Formula**" Pearl (2001):

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

Appealing result.

Identification result for natural indirect and direct effects

Under certain conditions (strong parametric assumptions, linear models, and no exposure-mediator interaction):

Resulting estimators same as Baron and Kenny (1986).

$\Rightarrow$ Causal literature adds a causal interpretation to those estimators.

Also: causal literature stresses the importance of including common causes C of M and Y, even in randomized settings, plus extends mediation analysis to different outcome types.

But: most causal methods need many counterfactual outcomes: not only $Y^{(1,M^{(0)})}$, also all $Y^{(a,m)}$: sometimes not conceivable.

$\Rightarrow$ Alternative: intervention-based approach

$\Rightarrow$ Intervention-based approach to mediation analysis:

Lok *Statistics In Medicine* (2016)
Lok and Bosch *Epidemiology* (2021)
https://arxiv.org/abs/1903.04697

No cross-worlds quantities, simple assumptions, product methods. Especially useful in identifying indirect effects of treatments to select promising treatments for further investigation in RCTs.

Interventions on the mediator

I: intervention on the mediator that does not affect pre-treatment common causes C of mediator M and outcome Y.

$M^{(a,I=1)}$ and $Y^{(a,I=1)}$: mediator and outcome under treatment a *and* under intervention I on the mediator.

"Organic" interventions on the mediator: Lok (2016), Lok and Bosch (2021)

Definition: (Organic intervention). An intervention I is an organic intervention relative to $a = 0$ and C if

$$M^{(0,I=1)} \mid C = c \quad \sim \quad M^{(1)} \mid C = c \tag{1}$$

and

$$Y^{(0,I=1)} \mid M^{(0,I=1)} = m, C = c \quad \sim \quad Y^{(0)} \mid M^{(0)} = m, C = c, \tag{2}$$

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

I is an organic intervention relative to $a = 1$ and C if (1) and (2) hold with the role of $a = 0$ and $a = 1$ reversed.

Organic intervention: interpretation

$$M^{(0,I=1)} \mid C = c \quad \sim \quad M^{(1)} \mid C = c$$

states that $I = 1$ "changes the distribution of the mediator to that under treatment", given C.

$$Y^{(0,I=1)} \mid M^{(0,I=1)} = m, C = c \quad \sim \quad Y^{(0)} \mid M^{(0)} = m, C = c,$$

states that $I = 1$ "has no direct effect on the outcome": after I affects the mediator, the system follows its natural course as under "no treatment", $a = 0$.

Organic intervention: common causes C

$$Y^{(0,I=1)} \mid M^{(0,I=1)} = m, C = c \quad \sim \quad Y^{(0)} \mid M^{(0)} = m, C = c,$$

Need all pre-treatment common causes of the mediator M and the outcome Y: C.

Without C: "mediator under treatment equals m" likely implies different prognosis:

under intervention I ($M^{(0,I=1)}$) versus without intervention ($M^{(0)}$),

since those convey different information on C, which in turn will lead to different expected outcomes.

"Organic" intervention: an example

Example: $A = 1$ blood pressure lowering medicine, M blood pressure, Y occurrence of a heart attack. Does $A = 1$ have a direct effect on heart attacks?

E.g., $A = 1$ lowers blood pressure by 10, on average for each C, without changing the shape of the blood pressure distribution. I intervention, in the untreated, that also decreases the blood pressure by 10, on average for each C, without changing the shape of the blood pressure distribution. $\Rightarrow M^{(0,I=1)} \sim M^{(1)}$ given C.

? I: Salt reduction in dosage dependent on $C = c$, perhaps 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).

$\Rightarrow$ Can hope

$$Y^{(0,I=1)} \mid M^{(0,I=1)} = m, C = c \quad \sim \quad Y^{(0)} \mid M^{(0)} = m, C = c$$

for this intervention I.

What does the salt reduction do?

- It keeps the effect mediated by the effect of A on blood pressure M: $M^{(0,I=1)} \sim M^{(1)}$ given C.
- It does not have the direct effect of $A = 1$.

"Organic" indirect and direct effects, Lok (2016):

$Y^{(a,I=1)}$: outcome under treatment a *and* under organic intervention I on the mediator.

Organic indirect effect of treatment A relative to $a = 0$ and C based on I:

$$EY^{(0,I=1)} - EY^{(0)}$$

((treatment= 0 for both $Y^{(0,I=1)}$ and $Y^{(0)}$, so mediated)).

Organic direct effect of treatment A relative to $a = 0$ and C based on I:

$$EY^{(1)} - EY^{(0,I=1)}$$

((mediator same distribution for $Y^{(1)}$ and $Y^{(0,I=1)}$, so not mediated)).

Interpretation of organic indirect effect

Interpretation of the organic indirect effect relative to $a = 0$:

The effect of an intervention that affects the mediator the same way $a = 1$ does, and then lets the system run its natural course as though there was no treatment and no intervention.

Organic versus natural indirect and direct effects

Organic indirect and direct effects relative to $a = 1$

generalize

Natural indirect and direct effects.

Provided that $M^{(1,I=1)} = M^{(0)}$ exists, the usual "cross-worlds" assumption implies that intervention I that sets the mediator to $M^{(0)}$, $M^{(1,I=1)} = M^{(0)}$, is organic!

Consistency assumption

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 $Y = Y^{(0)}$.

Identification result: Mediation Formula for organic indirect and direct effects relative to $a = 0$ and C

Under randomized treatment, estimation of $EY^{(0)}$ and $EY^{(1)}$ standard. For $E\left(Y^{(0,I=1)}\right)$:

Theorem *(Organic indirect and direct effects: the Mediation Formula for randomized data).* Under randomized treatment, consistency, and the definition of organic interventions, for an intervention I that is organic relative to $a = 0$ and C:

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

Note: does not depend on choice of organic intervention I!

Mediation formula for organic indirect and direct effects

Similar identification result as previously found for the natural indirect and direct effects studied by previous authors, of note Pearl (2001); Imai et al. (2010); Tchetgen-Tchetgen (2011).

Resulting expression in terms of observable quantities only. All three ingredients can be estimated using standard methods.

Our contribution: definition and therefore the interpretation of indirect and direct effects as well as the conditions under which estimators for these effects are valid can be considerably relaxed.

Organic indirect and direct effects: an intervention based approach

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

Organic indirect and direct effects and the product method in linear models

Combining organic interventions on the mediator with "treatment" (with $a = 1$), follows analogies with main stream natural indirect and direct effects.

Product method from Baron and Kenny (1986) works only in linear models when there is no treatment-mediator interaction in the outcome model.

Combining organic interventions on the mediator with "no treatment" (with $a = 0$, similar to "pure" indirect effect from Robins and Greenland (1992) see also Nguyen et al. (2020)):

Product method works in linear models regardless of treatment-mediator interaction in the outcome model.

Organic indirect and direct effects do not depend on choice of C

Theorem: If C has all pre-treatment common causes of the mediator M and the outcome $Y^{(a)}$, the organic indirect and direct effects relative to a are unique: which set of all pre-treatment common causes C does not affect the organic indirect and direct effects (Lok (2016)).

Thus, we can define:

Definition: When C includes all common causes of the mediator M and the outcome $Y^{(a)}$, we call the organic indirect and direct effects based on I relative to a and C *the* **organic indirect and direct effects relative to** a.

Selecting new treatments with promising indirect effects for clinical trials (!)

From the Mediation Formula:

Organic indirect effect relative to $a = 0$, $EY^{(0,I=1)} - EY^{(0)}$, can be estimated with:

(a) Distribution of mediator under "treatment", $A = 1$, and "no treatment", $A = 0$.

(b) Expectation of the outcome Y given the mediator M and pre-treatment covariates C under "no treatment", $A = 0$ (only under $A = 0$!).

This has important advantages in the pre-clinical stage, for selection of new treatments with promising indirect effects for clinical trials.

Relation with surrogate outcomes

Surrogate outcomes S, measured earlier than Y, are used in medical research to replace the outcome of interest Y, especially if it is thought that the effect of A on Y works mainly through the surrogate outcome S, or the surrogate outcome S is thought to reflect Y.

One could use mediation analysis to estimate the effect of A on Y if one knows the effect of A on S by considering S as a mediator. One could estimate the effect of A on Y by estimating the indirect effect of A on Y mediated by S. No on-treatment outcome data are needed for this.

Important lesson learnt from causal inference: we then need the common causes of S and Y.

Illustration: HIV cure: One could select HIV curative drugs which are at the pre-clinical stage for clinical trials, based on their estimated indirect effect

Most HIV-infected patients are on ART (recommended therapy). HIV viral load often undetectable on ART.

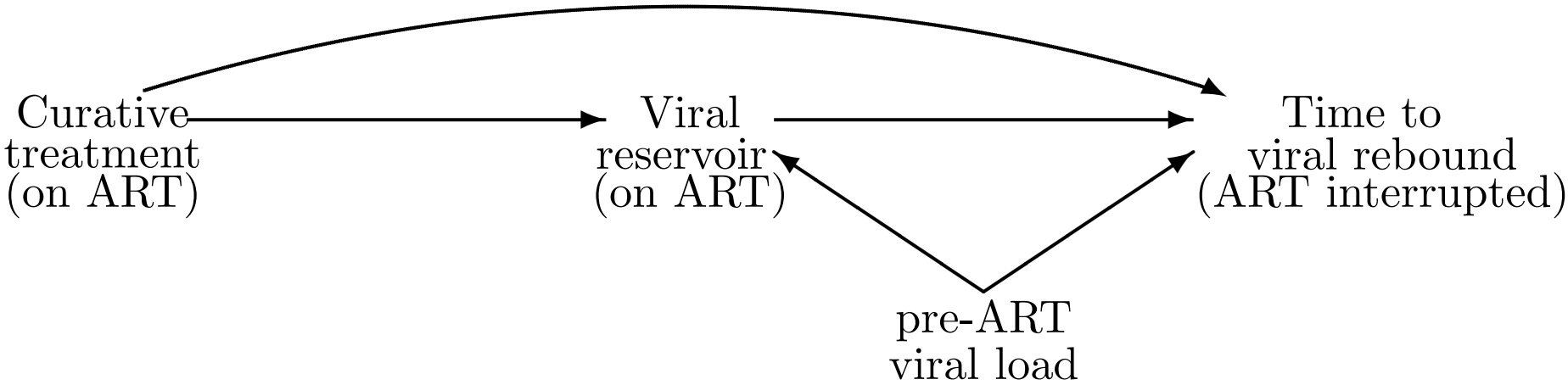
Aim of HIV cure: delay time to HIV viral rebound, the time until the HIV virus is detectable in the blood. Delay indefinitely? Not yet. Delay for some weeks...

HIV viral rebound not observable while on ART.

But: ART treatment interruption carries risks for the patients. Also expensive: close monitoring of patients off ART.

On-ART HIV viral reservoir measures are potential mediators M of the effect of HIV curative treatments A on viral rebound Y.

Figure 2: Causal diagram summarizing the HIV cure setting



Selecting HIV cure treatments with promising indirect effects for clinical trials (!)

ART treatment interruption has been carried out in a substantial number of patients in the absence of additional curative treatments ($A = 0$).

To estimate the *indirect* effect of the HIV curative treatments: Can combine the effect of HIV curative treatments ($A = 1$) on the on-ART viral reservoir M with ART interruption data under "no curative treatment" ($A = 0$).

Then **select potential HIV curative treatments with most promising indirect effects** for ART interruption trials.

The organic *indirect* effect of a curative HIV drug which affects a mediator

Outcome Y: suppressed ($Y = 1$)/not suppressed ($Y = 0$) by week 8 of ART interruption.

Two potential on-ART mediators M, both measures of the viral reservoir: cell associated HIV RNA and single-copy plasma HIV RNA, both on the $\log_{10}$ scale.

Pre-ART viral load is a measure of the size of the viral reservoir before ART initiation, and predictive of the time to viral rebound. No information on the pre-ART viral load $\Rightarrow$ the pre-ART NADIR CD4 count is a surrogate, our C.

The organic *indirect* effect of a curative HIV drug which affects a mediator

We estimate the organic indirect effect relative to $a = 0$ of a new curative HIV drug A, that is, the effect of a new curative HIV drug without a direct effect, that shifts the distribution of the mediator M by 1 $\log_{10}$ given C:

$$M^{(1)} \sim M^{(0)} - 1 \log_{10} \mid C. \tag{3}$$

Note 1: This is a shift in the *distribution* of the mediator rather than a shift of the actual values (no "rank preservation").

Note 2: We can estimate the *distribution* of $M^{(1)}$ under this shift by subtracting 1 $\log_{10}$ from the mediator $M^{(0)}$ under $a = 0$.

The organic *indirect* effect of a curative HIV drug which affects a mediator

Data: ACTG data from Li et al. (2016): data on 124 HIV infected patients without curative treaatments ($A = 0$), with on-ART HIV reservoir measures M, and with ART interruptions and viral rebound measurements Y.

Table 1: Organic indirect effects of curative HIV-treatments that shift the distribution of the reservoir measures downwards. Probability of virologic suppression without treatment was 63/124 or 51% (week 4) and 17/122 or 14% (week 8).

Mediator	Shift ($\log_{10}$ scale)[a]	Week[b,c]	Indirect effect[d]	95% CI[e]	RR[f]	95% CI[e]
SCA HIV-RNA[g]	0.5 $\log_{10}$	4	2.5%	(-2.5%,7.0%)	1.05	(0.96,1,13)
SCA HIV-RNA[g]	1 $\log_{10}$	4	5.7%	(-1.3%,12.6%)	1.10	(0.98,1.24)
SCA HIV-RNA[g]	∞[i]	4	6.2%	(-1.4%,14.0%)	1.12	(0.97,1.27)
CA HIV-RNA[h]	0.5 $\log_{10}$	4	6.9%	(1.7%,12.6%)	1.14	(1.03,1.26)
CA HIV-RNA[h]	1 $\log_{10}$	4	9.8%	(2.7%,17.0%)	1.19	(1.05,1.36)
CA HIV-RNA[h]	∞[i]	4	12.7%	(3.0%,22.7%)	1.25	(1.06,1.47)
SCA HIV-RNA[g]	0.5 $\log_{10}$	8	2.9%	(0.61%,6.3%)	1.18	(1.04,1.40)
SCA HIV-RNA[g]	1 $\log_{10}$	8	5.1%	(0.92%,10.1%)	1.32	(1.07,1.62)
SCA HIV-RNA[g]	∞[i]	8	5.7%	(0.95%,11.1%)	1.35	(1.07,1.68)
CA HIV-RNA[h]	0.5 $\log_{10}$	8	9.1%	(3.5%,15.5%)	1.66	(1.29,2.15)
CA HIV-RNA[h]	1 $\log_{10}$	8	11.9%	(4.9%,19.4%)	1.85	(1.40,2.45)
CA HIV-RNA[h]	∞[i]	8	15.9%	(6.8%,25.5%)	2.14	(1.56,2.92)

[a] Downwards shift of the mediator distribution, given the common cause C[d,e], on the $\log_{10}$ scale.
[b] For the week-4 analysis, analyzing viral rebound in the first 4 weeks after ART treatment interruption, C is NNRTI-based (versus not).
[c] For the week-8 analysis, analyzing viral rebound in the first 8 weeks after ART treatment interruption, C is the nadir CD4 count (categorized as ≤ 500 versus > 500). Not available in 2 patients, who were excluded from the week-8 analysis.
[d] Difference in probability of virologic suppression that is mediated.
[e] 95% Confidence Interval, calculated by bootstrap using Efron's percentile method (Van der Vaart 1998) with 5000 replicates. This leads to consistent coverage because of the Bootstrap Master Theorem (Kosorok 2008).
[f] Indirect effect effect on the Risk Ratio scale.
[g] Single-copy plasma HIV-RNA, on ART. Analysis restricted to the 94 patients with SCA HIV-RNA measured.
[h] Cell-associated HIV-RNA, on ART. All 124 patients had CA HIV-RNA measured.
[i] A which causes all mediator values below the limit of detection.

What did we obtain?

The organic *indirect* effect relative to $a = 0$ of a curative HIV drug that shifts the distribution of the viral reservoir by 1 $\log_{10}$ given pre-treatment common causes C, estimated from *existing ART interruption data under no curative treatment* ($A = 0$).

That is, *the effect of a curative HIV drug* that

- Shifts the distribution of the viral reservoir by 1 $\log_{10}$ given C
- And that *has no direct effect* on viral rebound.

Conclusions

Both for natural indirect and direct effects and for randomized indirect and direct effects Didelez et al. (2006): need to be able to **set** mediator to any specific value.

"Organic" indirect and direct effects: need to be able to affect the **distribution** of the mediator.

Organic indirect and direct effects relative to $a = 1$ generalize natural indirect and direct effects.

Conclusions

To answer clinical questions, often useful to combine interventions on the mediator with "no treatment", rather than with "treatment". Analogous to pure indirect effect of Robins and Greenland (1992).

Organic indirect effects relative to $a = 0$ can be estimated with information on

1. Distribution of mediator under treatment, $a = 1$, and "no treatment", $a = 0$.
2. Expectation of the outcome given the mediator and pre-treatment covariates C under $a = 0$ (only under $a = 0$!).

Important advantages in the pre-clinical stage, for selection of new treatments with promising indirect effects for clinical trials.

Closer to Baron and Kenny (1986).

Future and ongoing research

1. Inclusion of measurement error on the mediator. Approaches of Valeri and VanderWeele (2013) could be used?
2. What if the outcome logistic regression model also holds for mediators below the assay limit? Chernofsky et al. (2021), (Boston University Biostatistics PhD student).
3. Inclusion of post-treatment common causes of the mediator and the outcome; see also Lok (2020).
4. Other applications.
5. Do organic indirect and direct effects generalize separable indirect and direct effects?

Thanks for listening!

Acknowledgements organic indirect and direct effects

This research was supported by: NSF DMS 1854934, NIH/NIAID UM1 AI068634 and AI068636 (but not the view of NSF or NIH).

We thank anonymous referees for encouraging us to prove uniqueness of the organic indirect and direct effects, to define *the* organic indirect and direct effect as any version where C has all common causes, and for various other useful suggestions. And Ari Chernofsky (PhD student at the Department of Biostatistics, Boston University) for re-programming the point estimates for our HIV application, both for binary mediators and for mediator shifts, in R, in order to validate our SAS code.

We are extremely grateful to the HIV-infected participants who volunteered for the ACTG ART interruption trials. The authors also thank Dr. Li (Brigham and Women Hospital, Boston) for his support, for leading the treatment interruption project and generating the mediator data used in our analyses.

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.

Chernofsky, A., R. J. Bosch, and J. J. Lok (2021). Causal mediation analysis with mediator values below an assay limit. *arXiv preprint arXiv:2107.14782*.

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.

Didelez, 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.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science 25*(1), 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.

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. (2020). Organic direct and indirect effects with post-treatment common causes of mediator and outcome. *arXiv preprint arXiv:1510.02753*.

Lok, J. J. and R. J. Bosch (2021). Causal organic indirect and direct effects: Closer to the original approach to mediation analysis, with a product method for binary mediators. *Epidemiology 32*(3), 412–420. arXiv preprint arXiv:1903.04697.

Nguyen, T. Q., I. Schmid, and E. A. Stuart (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*. Also on https://arXiv.org/abs/1904.08515.

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 S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*, 143–155.

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.

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

Valeri, L. and T. J. VanderWeele (2013). Mediation analysis allowing for exposure-mediator interaction and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods 18*(2), 137–150.

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

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

The organic *indirect* effect of a curative HIV drug which affects a mediator

The organic indirect effect of a curative HIV drug that shifts the distribution of the mediator by 1 $\log_{10}$ given C equals

$$\begin{aligned}
&\int_{(m,c)} P\left(Y=1|M=m,C=c,A=0\right) f_{M|C=c,A=1}(m) f_C(c) dm\, dc \\
&\qquad -P\left(Y^{(0)}=1\right) \\
= &\int_{(m,c)} P\left(Y=1|M=\tilde{m}(m),C=c,A=0\right) f_{M|C=c,A=0}(m) f_C(c) dm \\
&\qquad -P\left(Y^{(0)}=1\right),
\end{aligned}$$

where $\tilde{m}(m) = m - 1\log_{10}$ if $m - 1\log_{10}$ is above the assay limit, and $\tilde{m}(m)$ is "below the assay limit" otherwise.

The organic *indirect* effect of a curative HIV drug which affects a mediator

We fit the following logistic regression model for the probability of virologic suppression at week 8 of ART interruption, Y, given M and C, using the data (y_i, m_i, c_i), $i = 1, \ldots, n$, without HIV curative drug ($A = 0$):

$$\begin{aligned} \text{logit} p_{\beta;m,c} &= \text{logit} P\left(Y_i = 1 | M_i = m, C_i = c, A_i = 0\right) \\ &= \begin{cases} \beta_0 + \beta_1 m + \beta_2 c & \text{if } m \text{ above the assay limit} \\ \beta_3 + \beta_2 c & \text{if } m \text{ below the assay limit.} \end{cases} \end{aligned}$$

The organic *indirect* effect of a curative HIV drug which affects a mediator

$$\int_{(m,c)} P\left(Y=1|M=m,C=c,A=0\right) f_{M|C=c,A=1}(m) f_C(c) dm\, dc$$

$$= \int_{(m,c)} P\left(Y=1|M=m,C=c,A=0\right) f_{M|C=c,A=0}(m+1\log_{10}) f_C($$

$$= \int_{(m,c)} P\left(Y=1|M=\tilde{m}-1\log_{10},C=c,A=0\right) f_{M|C=c,A=0}(\tilde{m}) f_C($$

In the last line we did a change of variables: $\tilde{m} = m + 1\log_{10}$.

Example of a binary mediator in HIV cure research

The estimated indirect effect (relative to $a = 0$) mediated by the binary mediator "below assay limit" (yes versus no).

For treatments that cause odds of "mediator below the assay limit" to increase by a factor 2, 3, or ∞ (∞: cause all mediator values to be below the assay limit; all: given C), we found:

Table 2: Organic indirect effects of curative HIV-treatments that increase the odds of reservoir measures below assay limit. Estimated probability of virologic suppression without treatment was 63/124 or 51% (week 4) and 17/122 or 14% (week 8).

Binary mediator	OR of below assay limit[a]	Week[b,c]	Indirect effect[d]	95% CI[e]	RR[f]	95% CI[e]
SCA HIV-RNA[g]	2	4	2.4%	(-0.5%,5.1%)	1.04	(0.99,1.10)
SCA HIV-RNA[g]	3	4	3.5%	(-0.7%,7.5%)	1.06	(0.99,1.14)
SCA HIV-RNA[g]	∞[i]	4	6.3%	(-1.3%,14.1%)	1.12	(0.98,1.27)
CA HIV-RNA[h]	2	4	3.7%	(0.86%,6.4%)	1.07	(1.02,1.13)
CA HIV-RNA[h]	3	4	5.7%	(1.3%,9.9%)	1.11	(1.03,1.21)
CA HIV-RNA[h]	10	4	10.0%	(2.3,17.5%)	1.20	(1.05,1.36)
CA HIV-RNA[h]	∞[i]	4	12.7%	(3.0%,22.7%)	1.25	(1.06,1.47)
SCA HIV-RNA[g]	2	8	2.2%	(0.4%,4.1%)	1.14	(1.03,1.25)
SCA HIV-RNA[g]	3	8	3.2%	(0.5%,6.0%)	1.20	(1.04,1.36)
SCA HIV-RNA[g]	∞[i]	8	5.7%	(0.9%,11.1%)	1.35	(1.07,1.68)
CA HIV-RNA[h]	2	8	4.0%	(1.7%,6.3%)	1.29	(1.15,1.43)
CA HIV-RNA[h]	3	8	6.4%	(2.8%,10.0%)	1.46	(1.23,1.69)
CA HIV-RNA[h]	10	8	11.9%	(5.1%,18.8%)	1.86	(1.43,2.36)
CA HIV-RNA[h]	∞[i]	8	15.9%	(6.9%,25.6%)	2.14	(1.56,2.92)

[a] Odds Ratio (OR) the mediator is below the assay limit, given the common cause $C^{b,c}$.
[b] For the week-4 analysis, analyzing viral rebound in the first 4 weeks after ART treatment interruption, C is NNRTI-based (versus not).
[c] For the week-8 analysis, analyzing viral rebound in the first 8 weeks after ART treatment interruption, C is the nadir CD4 count (categorized as ≤ 500 versus > 500). Not available in 2 patients, who were excluded from the week-8 analysis.
[d] Difference in probability of virologic suppression that is mediated by binary mediator.
[e] 95% Confidence Interval, calculated by bootstrap using Efron's percentile method (Van der Vaart 1998) with 5000 replicates. This leads to consistent coverage because of the Bootstrap Master Theorem (Kosorok 2008).
[f] Indirect effect effect on the Risk Ratio scale.
[g] Single-copy plasma HIV-RNA, on ART. Analysis restricted to the 94 patients with SCA HIV-RNA measured.
[h] Cell-associated HIV-RNA, on ART. All 124 patients had CA HIV-RNA measured.
[i] A which causes all mediator values below the limit of detection.

Definition of "C includes all common causes"

Definition: (*common cause*). X is *not* a common cause of mediator M and outcome $Y^{(a)}$ given C if either equation (4) or equation (5) holds:

$$X \perp\!\!\!\perp M^{(0)} \mid C \qquad \text{and} \qquad X \perp\!\!\!\perp M^{(1)} \mid C \tag{4}$$

or

$$X \perp\!\!\!\perp Y^{(a)} \mid M^{(a)}, C. \tag{5}$$

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

Product method for binary mediators: proof for organic indirect effect relative to $a = 0$

$$
\begin{aligned}
&\int_{(m,c)} E\left[Y|M=m, C=c, A=0\right]\left(f_{M|C=c,A=1}(m) - f_{M|C=c,A=0}(m)\right) f_C(c) dm\, dc \\
&= \int_c E\left[Y|M=1, C=c, A=0\right] \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc \\
&\quad + \int_c E\left[Y|M=0, C=c, A=0\right] \\
&\qquad \left(P(M=0|C=c, A=1) - P(M=0|C=c, A=0)\right) f_C(c)\, dc \\
&= \ldots - \int_c E\left[Y|M=0, C=c, A=0\right] \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc \\
&= \int_c \left(E\left[Y|M=1, C=c, A=0\right] - E\left[Y|M=0, C=c, A=0\right]\right) \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc.
\end{aligned}
$$

A product method for binary mediators

For binary mediators, under the conditions for the Mediation Formula, the organic indirect effect relative to $a = 0$ equals

$$\int_c (E[Y|M = 1, C = c, A = 0] - E[Y|M = 0, C = c, A = 0])$$
$$(P(M = 1|C = c, A = 1) - P(M = 1|C = c, A = 0)) f_C(c)\, dc;$$

a product method for binary mediators!

If A increases the probability of $M = 1$, it is the increased probability of $M = 1$ times the effect that changing $M = 0$ into $M = 1$ has on the outcome (under $a = 0$), averaged over C.

Product method for binary mediators: proof for organic indirect effect relative to $a = 1$

$$
\begin{aligned}
&\int_{(m,c)} E\left[Y|M=m, C=c, A=1\right]\left(f_{M|C=c,A=1}(m) - f_{M|C=c,A=0}(m)\right) f_C(c) dm\, dc \\
&= \int_c E\left[Y|M=1, C=c, A=1\right] \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc \\
&\quad + \int_c E\left[Y|M=0, C=c, A=1\right] \\
&\qquad \left(P(M=0|C=c, A=1) - P(M=0|C=c, A=0)\right) f_C(c)\, dc \\
&= \ldots - \int_c E\left[Y|M=0, C=c, A=1\right] \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc \\
&= \int_c \left(E\left[Y|M=1, C=c, A=1\right] - E\left[Y|M=0, C=c, A=1\right]\right) \\
&\qquad \left(P(M=1|C=c, A=1) - P(M=1|C=c, A=0)\right) f_C(c)\, dc.
\end{aligned}
$$

"Organic" indirect and direct effects relative to $a = 1$, Lok (2016):

Organic indirect and direct effects relative to $a = 1$ defined similarly, with role of $a = 0$ and $a = 1$ reversed.

$Y^{(a,I=1)}$: outcome under treatment a *and* under organic intervention I on the mediator.

Organic indirect effect of treatment A relative to $a = 1$ and C based on I:

$$EY^{(1)} - EY^{(1,I=1)}$$

((treatment= 1 for both $Y^{(1,I=1)}$ and $Y^{(1)}$, so mediated)).

Organic direct effect of treatment A relative to $a = 1$ and C based on I:

$$EY^{(1,I=1)} - EY^{(0)}$$

((mediator same distribution for $Y^{(1,I=1)}$ and $Y^{(0)}$, so not mediated)).

Interpretation of organic indirect effect relative to $a = 1$

Interpretation of the organic direct effect relative to $a = 1$:

The effect of an intervention that has the same effect through other pathways as $a = 1$, but leaves the mediator distribution as under $a = 0$.

The Mediation Formula for organic indirect and direct effects

Proof:

$$E\left(Y^{(0,I=1)}\right) = E\left(E\left[Y^{(0,I=1)}|M^{(0,I=1)}, C\right]\right)$$

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

because of the definition of conditional expectations,

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

because of the definition of organic intervention. □

Product method for organic indirect effect relative to $a = 0$ for linear models: assumptions

Assumption:

$$E[Y|M = m, A = 0, C = c] = \beta_0 + \beta_1 c + \beta_2 m.$$

Note: $E[Y|M = m, A = a, C = c]$ can have an interaction term $\beta_3 am$.

Assumption:

$$M^{(1)} \sim M^{(0)} + \alpha_1 \mid C,$$

i.e.

$$f_{M^{(1)}|C=c}(m) = f_{M^{(0)}|C=c}(m - \alpha_1).$$

$$
\begin{aligned}
& E\left(Y^{(0,I=1)}\right) - EY^{(0)} \\
= & \int_{(m,c)} E\left[Y|M=m, A=0, C=c\right] \\
& \quad \left(f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m)\right) f_C(c) dm\, dc \\
= & \int_{(m,c)} \left(\beta_0 + \beta_1 c + \beta_2 m\right) \left(f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m)\right) \\
& \quad f_C(c) dm\, dc \\
= & \beta_2 \int_{(m,c)} m \left(f_{M|A=1,C=c}(m) - f_{M|A=0,C=c}(m)\right) f_C(c) dm\, dc \\
= & \beta_2 \int_{(m,c)} m\, f_{M|A=0,C=c}(m-\alpha_1) f_C(c) dm\, dc \\
& -\beta_2 \int_{(m,c)} m\, f_{M|A=0,C=c}(m) f_C(c) dm\, dc \\
= & \beta_2 \left(\int_{(\tilde{m},c)} (\tilde{m}+\alpha_1) f_{M|C=c,A=0}(\tilde{m}) f_C(c) d\tilde{m}\, dc - \ldots \right) \\
= & \beta_2 \alpha_1
\end{aligned}
$$

Uniqueness of organic indirect and direct effect

Next: organic indirect and direct effect do not depend on C if C "has all common causes of mediator M and outcome Y".

Uniqueness of organic indirect and direct effects

Let I^C be an intervention that is organic with respect to C and $I^{\tilde{C}}$ and intervention that is organic with respect to $\tilde{C}$. Assume that C is not a common cause of mediator and outcome given $\tilde{C}$, and $\tilde{C}$ is not a common cause of mediator and outcome given C; hence there are 4 different cases, with either (4) or (5) holding for C and $\tilde{C}$, respectively. One can show show that under any of the 4 different cases,

$$E\left(Y^{(0,I^{\tilde{C}}=1)}\right) = \int_{(\tilde{c},m,c)} E\left[Y^{(0)} \mid M^{(0)} = m, \tilde{C} = \tilde{c}, C = c\right] f_{M^{(1)}|\tilde{C}=\tilde{c},C=}$$

Because of symmetry, it follows that also

$$E\left(Y^{(0,I^{C}=1)}\right) = \int_{(\tilde{c},m,c)} E\left[Y^{(0)} \mid M^{(0)} = m, \tilde{C} = \tilde{c}, C = c\right] f_{M^{(1)}|\tilde{C}=\tilde{c},C=}$$

But then, $E\left(Y^{(0,I^{\tilde{C}}=1)}\right) = E\left(Y^{(0,I^{C}=1)}\right)$.

Organic indirect and direct effects generalize natural indirect and direct effects

Theorem: If $M^{(1,I=1)} = M^{(0)}$ exists, identifiability assumption for natural indirect and direct effects,

$$Y^{(a',m)} \perp\!\!\!\perp M^{(a)} | C = c, A = a, \qquad (*)$$

implies that $M^{(1,I=1)} = M^{(0)}$ is organic intervention.

Proof : $?Y^{(1,I=1)}|M^{(1,I=1)} = m, C = c \sim Y^{(1)}|M^{(1)} = m, C = c?$

$$\Rightarrow ?Y^{(1,m)}|M^{(0)} = m, C = c \sim Y^{(1,m)}|M^{(1)} = m, C = c? \qquad (6)$$

And yes: under randomization, $A = a$ can be left out of the conditioning event of (*), which hence implies (6): given C, $Y^{(1,m)}$ depends neither on $M^{(0)}$ nor on $M^{(1)}$.

From natural conditions to organic conditions

If can envision all $Y^{(a,m)}$: my interpretation: nature did not have more information on the potential outcomes $Y^{(a',m)}$ to determine the value of the mediator $M^{(a)}$ than recorded in C.

For estimation, needs that all common causes of mediator and outcome are recorded in the database. Seems strong condition, but usual. Methods to estimate organic indirect and direct effects assume this as well.

A weaker identifiability condition for natural indirect and direct effects

For natural indirect and direct effect, consider $M^{(1,I=1)} = M^{(0)}$. If all $Y^{(a,m)}$ "exist", this I is an organic intervention if:

$$Y^{(1,m)} | M^{(0)} = m, C = c \quad \sim \quad Y^{(1,m)} | M^{(1)} = m, C = c.$$

((Follows e.g. if (*) holds)). Then, natural and organic indirect and direct effects are the same.

Still cross-worlds assumption. No surprise: definition cross-worlds.

Relaxes (*) though.

Comparison with Didelez et al. (2006): randomize mediator and then set

Didelez et al. (2006) propose to let $M^{(1,I=1)}$ be: take a random draw from $M^{(0)}$ given C, then set the mediator to this random draw. Such intervention I is organic if

$$Y^{(1,I=1)}|M^{(1,I=1)} = m, c = c \quad \sim \quad Y^{(1)}|M^{(1)} = m, C = c.$$

Can be tested: not cross-worlds. E.g. (*) also implies that this $M^{(1,I=1)}$ is organic. Then, Didelez et al. (2006) indirect and direct effects same as organic indirect and direct effects, provided one can set the mediator to a random draw.

Observational data

Same definition, similar identification result provided "C has all common causes of mediator and outcome".

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

Assumption: *(No Unmeasured Confounding).*

$$A \perp\!\!\!\perp \left(Y^{(0)}, M^{(0)} \right) \mid C, Z$$

and

$$A \perp\!\!\!\perp Y^{(1)} \mid C, Z \quad \text{and} \quad A \perp\!\!\!\perp M^{(1)} \mid C, Z.$$

Observational data: identification result

Theorem: *(Organic indirect and direct 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 mediator and outcome. Then

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

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

Mother-to-child transmission of HIV/AIDS

HIV-infected mothers can transmit the HIV virus to their infants.

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)).

Mother-to-child transmission of HIV/AIDS

???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???

Outcome Y: indicator "newborn baby is HIV-infected".

Mediator M: HIV-1 RNA.

I: 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.

Quantity of interest: $EY^{(0,I=1)} - EY^{(0)}$.

Note: intervention I on the mediator combined with "no treatment", more relevant here.

Mother-to-child transmission of HIV/AIDS

HIV viral load in the mother's blood cannot be "set".

Once we will be able to set it, we will set it to 0.

$\Rightarrow$ For a treatment like AZT, organic indirect and direct effects are more natural than their natural counterparts.

Mother-to-child transmission of HIV/AIDS

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.

Whether intervention is organic can (and should) be discussed with subject matter experts.