Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Web-appendix with: Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible

Judith J. Lok

B. Web-appendix: The smoking-and-low-birth-weight paradox

Section 4 argued that if a common cause of mediator and outcome $\tilde{C}$ has not been observed, it is often not reasonable to think that equation (4.4) without $\tilde{C}$ would hold. As an example, this appendix considers the case of maternal smoking and infant mortality. The effect of smoking during pregnancy ($A = 1$) on infant mortality may be mediated by low birth weight. It turns out that a naive analysis leads to the conclusion that the direct effect of maternal smoking on infant mortality is beneficial. [?] explain this "birth weight paradox" and provide an explanation for the possible biases. This appendix shows how this relates to the setup of this article.

For exposition simplicity assume that whether a pregnant woman smokes or not is unrelated to her prognosis with respect to low birth weight or complications in her infant in the "smoking" and "not smoking" scenarios. So, differences in outcomes between smokers and nonsmokers are caused by smoking only, effectively implying that the treatment "smoking" can be considered randomized. In practice, this may be violated if women with other unhealthy behaviors besides smoking are more likely to smoke. Those complications are ignored here, because the issues addressed in this appendix are present even under randomized treatment, and relaxing the randomization assumption was already discussed in Section 7.

Some infants may have a low birth weight due to genetically determined birth defects, which are likely not caused by smoking, or due to environmental causes other than smoking like malnutrition. These causes may be more predictive of infant mortality than smoking ([?]). For exposition simplicity this appendix bases the discussion on genetically determined birth defects as common causes of birth weight and infant mortality. Denote these by $\tilde{C}$. Suppose that, as in most studies, $\tilde{C}$ is not observed. Now consider an intervention I (Definition 4.1 equation (4.3)) which causes birth weight for the smoking mothers to have the same distribution as the birth weight for non-smoking mothers, without changing genetically determined birth defects $\tilde{C}$. Then, the prognosis of an infant had the mother smoked and "had the infant had a normal birth weight $M_{1,I=1}$ under intervention I" is most likely not the same as the prognosis of an infant had the mother smoked and "had the infant had normal birth weight M_1 without intervention". Without the intervention, in an infant of a smoking mother with normal birth weight M_1, genes responsible for birth defects are most likely more favorable: the birth weight was normal without intervention, even while the mother was smoking. So, one would think that the prognosis Y_1 is good for such an infant. Under intervention I, some of the infants of smoking women with normal birth weight $M_{1,I=1}$ will have genetically determined birth defects: the birth weight has been intervened on to be normal without changing genetically determined birth defects. The possibility of genetically determined birth defects would lead to a worse prognosis $Y_{1,I=1}$ for such infants. Thus, equation (4.4) will generally not hold in this situation.

Next, I consider how this issue affects the estimators of the direct and indirect effects if $\tilde{C}$ is ignored (which it has to be, because it is assumed that $\tilde{C}$ is unobserved). Let the outcome Y be an indicator of infant mortality, and let I be an intervention for which equation (4.3) holds. If $\tilde{C}$ is ignored, $E(Y_{1,I=1})$ would be estimated using the data for women who smoked but who had infants with relatively high birth weights, because that is the distribution of the birth weights $M_{1,I=1}$ under intervention I. As argued in the previous paragraph, this approach is too optimistic, and thus the mortality probability $E(Y_{1,I=1})$ is underestimated. Thus, the part of the effect of smoking that is mediated through low birth weight, the indirect effect of smoking, is overestimated. As a consequence, the direct effect of smoking on infant mortality is underestimated.

This is in line with what was found in e.g. [?], who studied controlled direct effects, and found that conditional on birth weight, smoking and infant mortality were negatively associated in infants with low birth weight. A naive approach would thus conclude that the direct effect of smoking is beneficial. [?] explained this by noting that low birth weight may be more harmful if caused by genetic birth defects than if caused by smoking. As outlined above, this is a violation of equation (4.4).

The solution to this issue is to try to include in C as many pre-treatment common causes of mediator and outcome as feasible. In the case of the genetically determined birth defects in the above example, this could perhaps be done through observed traits of the newborn babies. If this is unfeasible, conclusions may be flawed because equation (4.4) fails to hold. The direction of the bias can be reasoned as described in the previous paragraph: in this example, ignoring birth defects results in an overestimation of the organic indirect effect of smoking (mediated by birth weight), and an underestimation of the detrimental organic direct effect of smoking (not mediated by birth weight).

The discussion in this section illustrates the importance of the assumptions behind mediation analysis. One can compare whether the distribution on the left hand side of equation (4.4) puts more mass on larger values of the outcome or on smaller values of the outcome as compared to the distribution on the right hand side. Thus, an advantage of the current approach is that the direction of the bias that results from lack of validity of equation (4.4) can be discussed in the context of each particular application.

C. Web-appendix: Interventions on the mediator under treatment or on the mediator under no treatment?

There has been some discussion in the previous literature about whether one should consider setting the mediator to its value under treatment versus setting it to its value without treatment (see e.g. [?]). As indicated in Section 8, the approach in this article can easily be extended to incorporate both. To illustrate what might be of most clinical interest in a particular setting, consider two scenarios. In scenario 1, an alternative treatment I' changes the mediator the same way conventional treatment A does, without having a direct effect on the outcome. This would be especially relevant for example if the direct effect of treatment A is a harmful side effect. In this case, one would want to compare the distribution of the outcome under no treatment with the distribution of the outcome under no treatment if the mediator under intervention I', $M_{0,I'=1}$, has the same distribution as M_1. For example, with $Y_{0,I'=1}$ the outcome under I' with $A=0$, one would want to estimate $E\big(Y_{0,I'=1}-Y_0\big)$ as the effect mediated through M. This is a different quantity than the organic indirect effect of Section 4, but can be estimated in a similar way by changing the coding of A as in Section 8. In scenario 2, the quantity of interest is the effect of an alternative treatment $\tilde{A}$, where $\tilde{A}$ has the same direct effect as treatment A, but does not affect the mediator M. This would be especially relevant if the effect of treatment A on the mediator is a harmful side effect. In this case one would want to consider an intervention such that $M_{1,I=1}$ under treatment has the same distribution as M_0. In that situation, the quantity of interest is $E\big(Y_{1,I=1}-Y_0\big)$, the organic direct effect of treatment $A=1$ as defined in Definition 4.1. Scenario 1 motivates an intervention that causes the mediator without treatment to have the same distribution as M_1, scenario 2 motivates an intervention that causes the mediator with treatment to have the same distribution as M_0. When studying the biological mechanisms by which particular treatments are effective, both types of interventions may be of interest.

D. Web-appendix: Inference under randomized treatment

I now illustrate how one might use the identification result of Section 6 to estimate $E\left(Y_{1,I=1}\right)$, and hence the organic indirect and direct effects, under semi-parametric assumptions.

Suppose that $M_1 \sim M_0 + \beta_1 + \beta_3^\top C \mid C$, with $\beta_1 \in \mathbb{R}$ and $\beta_3 \in \mathbb{R}^k$. This would be the case if, as in e.g. [?], M follows a regression model $M = \beta_0 + \beta_1 A + \beta_2^\top C + \beta_3^\top AC + \epsilon$, where the random variable ϵ has the same distribution given C under treatment as without treatment, and with $\beta_1 \in \mathbb{R}$ and $\beta_2, \beta_3 \in \mathbb{R}^k$. Suppose in addition that the expected value of Y given C and M under treatment follows some parametric model of the form $E\left[Y \mid M=m, C=c, A=1\right] = f_\theta(m,c)$. Notice that this last model applies only to the distribution of Y conditional on $A=1$, not conditional on $A=0$. This implies that the model does not restrict treatment-mediator interactions. Then, Theorem 6.1 implies $E\left(Y_{1,I=1}\right)$

$= E\left[f_\theta(M - \beta_1 - \beta_3^\top C, C) \mid A = 1\right]$ (proof: see below). This can be estimated by fitting the models for β and θ using standard methods, plugging the parameter estimates in, and replacing the expectation given $A = 1$ by its empirical average. Standard errors can be estimated with the bootstrap.

Notice that the resulting estimator uses changes in the distribution of the mediator with and without treatment, but the distribution of the outcome only in treated units. This leads to an estimator for the indirect effect that does not use data on the outcomes for untreated units.

[?] provide code to estimate direct and indirect effects based on the mediation formula for the case where M and Y both follow regression or logistic regression models.

Proof of inference under randomized treatment:

$$\begin{aligned}
E\left(Y_{1,I=1}\right) &= \int_{(c,m)} E\left[Y \mid M = m, C = c, A = 1\right] f_{M|C=c,A=0}(m) f_C(c) dm dc \\
&= \int_{(c,m)} f_\theta(m, c) f_{M|C=c,A=1}(m + \beta_1 + \beta_3^\top c) f_C(c) dm dc \\
&= \int_{(c,\tilde{m})} f_\theta(\tilde{m} - \beta_1 - \beta_3^\top c, c) f_{M|C=c,A=1}(\tilde{m}) f_C(c) d\tilde{m} dc \\
&= E\left[f_\theta(M - \beta_1 - \beta_3^\top C, C) \mid A = 1\right],
\end{aligned}$$

where the first equality follows from Theorem 6.1, the second equality follows from $M_1 \sim M_0 + \beta_1 + \beta_3^\top C \mid C$, see above, the third equality from a change of variables with $\tilde{m} = m + \beta_1 + \beta_3^\top c$, and the fourth equality from the fact that treatment A is randomized, and therefore the distribution of C does not depend on A. □

E. Web-appendix: Organic direct and indirect effects: independence assumptions instead of distributional assumptions

Some readers may be more at ease with independence assumptions underlying causal inference than with the distributional assumptions considered in the main text. This can be done in the current context as follows. Let R describe the possible treatments as follows: $R = 0$: treatment 0, $R = 1$: treatment 1 and $R = 2$: treatment 1 combined with an "organic" intervention I on the mediator. Equivalent to the definition in the main text, the definition for I being an organic intervention on the mediator could be formulated as that both equations (1) and (2) are satisfied:

$$M \perp\!\!\!\perp R \mid C = c, R \neq 1 \tag{1}$$

$$Y \perp\!\!\!\perp R \mid M = m, C = c, R \neq 0. \tag{2}$$

Of course, for easier interpretation, $R \neq 1$ could be replaced by "$R = 0$ or $R = 2$" and $R \neq 0$ could be replaced by "$R = 1$ or $R = 2$". The first of these assumptions states that, for given pre-treatment covariates C, the mediator is independent of whether the mediator was intervened on during treatment versus no treatment was given. The second of these assumptions states that, for given mediator and pre-treatment covariates C, the outcome is independent of whether the mediator got its value m because it was intervened on during treatment versus treatment 1 was given.

F. Web-appendix: Organic direct and indirect effects without counterfactuals

Some of the literature on causal inference is avoiding counterfactuals, see e.g. [?], [?], and [?]. Although this has not been a concern in the main manuscript, some readers may appreciate that organic direct and indirect effects can also be defined without counterfactuals, if "organic" interventions are possible in a three-arm clinical trial with $R = 0$: treatment 0, $R = 1$: treatment 1 and $R = 2$: treatment 1 combined with an "organic" intervention I on the mediator. In this setting, the definition for I being an organic intervention on the mediator is that both equations (3) and (4) are satisfied:

$$M \mid R = 2, C = c \sim M \mid R = 0, C = c \tag{3}$$

$$Y \mid R = 2, M = m, C = c \sim Y \mid R = 1, M = m, C = c. \tag{4}$$

Equation (3) states that the distribution of the mediator under treatment combined with the intervention I is as under treatment 0, and equation (4) intuitively states that the intervention I on the mediator has no direct effect on the outcome Y.

The organic direct and indirect effects based on I can now be defined as

$$E[Y|R=1]-E[Y|R=2]$$

and

$$E[Y|R=2]-E[Y|R=0].$$

As in the main paper, the mediation formula holds for $E[Y \mid R=2]$ because

$$\begin{aligned} E[Y \mid R=2] &= E(E[Y|M,C,R=2]) \\ &= \int_{m,c} E[Y|M=m,C=c,R=2] f_{M|C=c,R=2}(m) f_{C|R=2}(c) \\ &= \int_{m,c} E[Y|M=m,C=c,R=1] f_{M|C=c,R=0}(m) f_{C}(c), \end{aligned}$$

because R is randomized, (3), and (4).