Online Supplemental Materials A

Interventional indirect effects when there are more than two mediators

Definitions

In general for $t > 2$ mediators, the indirect effect via each distinct mediator $M_s, s = 1, \ldots, t$, can be defined as:

$$\text{IE}_s = \text{E}\left[\sum_{m_1,\ldots,m_t} \text{E}(Y_{1m_1\cdots m_t}|C)\Big\{\Pr(M_{s,1} = m_s|C) - \Pr(M_{s,0} = m_s|C)\Big\} \times \prod_{k=1}^{s-1} \Pr(M_{k,1} = m_k|C) \prod_{l=s+1}^{t} \Pr(M_{l,0} = m_l|C)\right], \quad \text{(SI.1)}$$

where the individual values of the first $s-1$ mediators are randomly drawn from their respective (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level(s) $a^{(1)} = \ldots = a^{(s-1)} = 1$, and the last $t-s$ mediators are randomly drawn from their respective (counterfactual) marginal distribution (given covariates C) under hypothetical treatment level(s) $a^{(s+1)} = \ldots = a^{(t)} = 0$. The interventional indirect effect of treatment on outcome via the mediator M_s is interpreted as the combined effect along all (underlying) causal pathways from A to M_s (possibly intersecting any other mediators that are causes of M_s), then lead directly from M_s to Y.

Note that the proposed definition in (SI.1) differs from the existing definition in Vansteelandt & Daniel (2017, eAppendix B, Equation (1)): under the latter definition, the mediators are drawn from a mixture of joint, and possibly marginal, distributions. Under the definition in (SI.1), each mediator is always drawn from its marginal distribution, hence avoiding joint distributions of different subsets of the (other) mediators when assessing shifts in the marginal distribution for a particular mediator. The proposed definition in (SI.1) is intended to make a clearer distinction between the definitions of the joint (for all mediators) and separate (via each mediator) indirect effects, thus permitting a simpler expression for the indirect effect via the mediators' mutual dependence as defined in (SI.4) below. Because the proposed and existing definitions target fundamentally different causal effects, comparisons of their substantive relevance under different research settings, or the finite sample behaviors of

their respective estimators (under certain settings where the effects may be equivalent), are beyond the scope of this article and deferred to future work.

The sum of the separate indirect effects via each mediator is therefore:

$$\mathrm{E}\left[\sum_{m_1,\ldots,m_t}\mathrm{E}(Y_{1m_1\cdots m_t}|C)\left\{\prod_{s=1}^{t}\Pr(M_{s,1}=m_s|C)-\prod_{s=1}^{t}\Pr(M_{s,0}=m_s|C)\right\}\right]; \tag{SI.2}$$

the joint indirect effect via the mediators is similarly defined as:

$$\mathrm{E}\Bigg[\sum_{m_1,\ldots,m_t}\mathrm{E}(Y_{1m_1\cdots m_t}|C)\left\{\Pr(M_{1,1}=m_1,\ldots,M_{t,1}=m_t|C)\right.$$
$$\left.-\Pr(M_{1,0}=m_1,\ldots,M_{t,0}=m_t|C)\right\}\Bigg]. \tag{SI.3}$$

The difference between (SI.3) and (SI.2) is defined to be the interventional indirect effect due to the mediators' mutual dependence on one another:

$$\mathrm{E}\Bigg[\sum_{m_1,\ldots,m_t}\mathrm{E}(Y_{1m_1\cdots m_t}|C)\Bigg\{\Pr(M_{1,1}=m_1,\ldots,M_{t,1}=m_t|C)-\prod_{s=1}^{t}\Pr(M_{s,1}=m_s|C)$$
$$-\Pr(M_{1,0}=m_1,\ldots,M_{t,0}=m_t|C)+\prod_{s=1}^{t}\Pr(M_{s,0}=m_s|C)\Bigg\}\Bigg]. \tag{SI.4}$$

The interventional direct effect of treatment on outcome that avoids all t mediators is correspondingly defined as:

$$\mathrm{E}\left[\sum_{m_1,\ldots,m_t}\left\{\mathrm{E}(Y_{1m_1\cdots m_t}|C)-\mathrm{E}(Y_{0m_1\cdots m_t}|C)\right\}\Pr(M_{1,0}=m_1,\ldots,M_{t,0}=m_t|C)\right]. \tag{SI.5}$$

Estimation

In the following, we derive closed form expressions for the estimators of the interventional (in)direct effects defined above. We will assume (correct) models for the mean outcome and joint distribution of the mediators, and that assumptions (A1)–(A3) in the main text are met.

Outcome models without interactions. We first derive estimators of the interventional (in)direct effects via each mediator assuming a linear and additive mean model for the outcome; i.e.,

$$\mathrm{E}(Y|A,M_1,\ldots,M_t,C)=\beta_0+\beta_A A+\sum_{s=1}^{t}\beta_s M_s+\beta_C C. \tag{SI.6}$$

Suppose that the overall treatment effect on each mediator, given baseline covariate(s) C, is parametrized by the (partial) regression coefficient of treatment A in the following linear and additive (marginal) mean model for each mediator:

$$\mathrm{E}(M_s|A,C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, \quad s = 1, \ldots, t. \tag{SI.7}$$

The outcome mean model and mediator (marginal) mean models for the setting with $t = 2$ mediators are stated as (1) and (2) in the main text respectively. It follows that the interventional direct effect and indirect effects via each mediator $M_s, s = 1, 2$, are identified by (functions of) the parameters in the assumed models; i.e., $\mathrm{DE} = \beta_A$ and $\mathrm{IE}_s = \beta_s \delta_s$ respectively. Again we note that the overall effect of A on M_s, as encoded by δ_s in (SI.7), captures all of the underlying treatment effects that are transmitted through any other mediators that are causes of M_s.

As previously noted, the mean models (SI.6) and (SI.7) adopt the same functional form for the expected values of the outcome Y and mediators $M_s, s = 1, 2$, as a parallel path model where the mediators are assumed not to causally affect one another. The interventional indirect effect via each mediator M_s (SI.1) equals the indirect effect using the product-of-coefficients method $\beta_s \delta_s$ for the path $A \to M_s \to Y$ in the parallel path model. Similarly, the interventional direct effect (SI.5) equals the path coefficient β_A for the path $A \to Y$ that avoids all the mediators in the parallel path model. However, the assumed mean model for the outcome in (SI.6) implies a zero indirect effect via the mediators' mutual dependence on one another (SI.4), because the joint indirect effect via all mediators (SI.3) equalled the sum of the separate indirect effects via each mediator (SI.2). Estimators of the interventional effects can thus be obtained by fitting the parallel path model to the observed data using SEM or OLS, then plugging in estimates of the (partial) regression coefficients in the respective direct and indirect effects. Standard errors can be estimated using a nonparametric percentile bootstrap procedure (Efron & Tibshirani, 1994) that randomly resamples observations with replacement.

Outcome models with treatment-mediator, mediator-mediator, and treatment-mediator-mediator interactions. We now derive estimators of the interventional (in)direct effects when the outcome model includes treatment-mediator,

mediator-mediator, and treatment-mediator-mediator interactions; i.e.,

$$\mathrm{E}(Y|A, M_1, \ldots, M_t, C) = \beta_0 + \beta_A A + \sum_{s=1}^{t}(\beta_s + \beta_{As}A)M_s + \sum_{\substack{k,l=1,\\ k<l}}^{t}(\beta_{kl} + \beta_{Akl}A)M_k M_l + \beta_C C. \tag{SI.8}$$

The interventional indirect effect via each mediator $M_s, s = 1, \ldots, t$, in (SI.1) is identified by:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m_1,\ldots,m_t} \mathrm{E}(Y|1, m_1, \ldots, m_t, C)\Big\{\Pr(M_s = m_s|A = 1, C) - \Pr(M_s = m_s|A = 0, C)\Big\}\right. \\
&\quad \left.\times \prod_{k=1}^{s-1}\Pr(M_k = m_k|A = 1, C) \prod_{l=s+1}^{t}\Pr(M_l = m_l|A = 0, C)\right] \\
&= \left\{(\beta_s + \beta_{As}) + \sum_{k=1}^{s-1}(\beta_{ks} + \beta_{Aks})\,\mathrm{E}(M_k|A = 1) + \sum_{l=s+1}^{t}(\beta_{sl} + \beta_{Asl})\,\mathrm{E}(M_l|A = 0)\right\}\delta_s.
\end{aligned} \tag{SI.9}$$

The sum of the separate indirect effects via each mediator (SI.2) is thus identified by:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m_1,\ldots,m_t} \mathrm{E}(Y|1, m_1, \ldots, m_t, C)\left\{\prod_{s=1}^{t}\Pr(M_s = m_s|A = 1, C) - \prod_{s=1}^{t}\Pr(M_s = m_s|A = 0, C)\right\}\right. \\
&= \sum_{s=1}^{t}(\beta_s + \beta_{As})\delta_s + \sum_{\substack{k,l=1,\\ k<l}}^{t}(\beta_{kl} + \beta_{Akl})\left\{\mathrm{E}(M_k|A = 1)\,\mathrm{E}(M_l|A = 1) - \mathrm{E}(M_k|A = 0)\,\mathrm{E}(M_l|A = 0)\right\}.
\end{aligned}$$

Whereas the joint indirect effect due to all t mediators is identified by:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m_1,\ldots,m_t} \mathrm{E}(Y|1, m_1, \ldots, m_t, C)\right. \\
&\quad \left.\times\left\{\Pr(M_1 = m_1, \ldots, M_t = m_t|A = 1, C) - \Pr(M_1 = m_1, \ldots, M_t = m_t|A = 0, C)\right\}\right] \\
&= \sum_{s=1}^{t}(\beta_s + \beta_{As})\delta_s + \sum_{\substack{k,l=1,\\ k<l}}^{t}(\beta_{kl} + \beta_{Akl})\{\mathrm{E}(M_k M_l|A = 1) - \mathrm{E}(M_k M_l|A = 0)\}.
\end{aligned}$$

Denote the $t \times t$ covariance matrix of the mediators that depends on treatment A by $\Sigma(A)$, with the (k, l) entry being $\Sigma_{kl}(A), k, l = 1, \ldots, t$. Under the outcome mean model (SI.8) and marginal mean models (SI.7) for each mediator, the indirect effect due to the

mediators' mutual dependence is:

$$
\begin{aligned}
&\mathrm{E}\Bigg[\sum_{m_1,\ldots,m_t} \mathrm{E}(Y|A=1,m_1,\ldots,m_t,C) \\
&\quad\times\Big\{\Pr(M_1=m_1,\ldots,M_t=m_t|A=1,C)-\prod_{s=1}^{t}\Pr(M_s=m_s|A=1,C) \\
&\qquad-\Pr(M_1=m_1,\ldots,M_t=m_t|A=0,C)+\prod_{s=1}^{t}\Pr(M_s=m_s|A=0,C)\Big\}\Bigg] \\
&=\sum_{\substack{k,l=1,\\ k<l}}^{t}(\beta_{kl}+\beta_{Akl})\,\mathrm{E}\{\mathrm{cov}(M_k,M_l|A=1,C)-\mathrm{cov}(M_k,M_l|A=0,C)\} \\
&=\sum_{\substack{k,l=1,\\ k<l}}^{t}(\beta_{kl}+\beta_{Akl})\{\Sigma_{kl}(1)-\Sigma_{kl}(0)\}. \qquad \text{(SI.10)}
\end{aligned}
$$

The indirect effect (SI.10) is non-zero when there are mediators whose (i) interaction affects the outcome ($\beta_{kl}+\beta_{Akl}\neq 0$) , and (ii) covariance is affected by treatment ($\Sigma_{kl}(1)-\Sigma_{kl}(0)\neq 0$). Furthermore, the indirect effect is the (weighted) sum of the treatment effects on each element of the covariance matrix for the mediators.

The joint indirect effect is exactly the sum of the indirect effects via each mediator, and via the mediators' mutual dependence. This follows from the identity:

$$
\begin{aligned}
&\mathrm{E}(M_kM_l|A=1)-\mathrm{E}(M_kM_l|A=0) \\
&=\{\mathrm{cov}(M_k,M_l|A=1)+\mathrm{E}(M_k|A=1)\,\mathrm{E}(M_l|A=1)\} \\
&\quad-\{\mathrm{cov}(M_k,M_l|A=0)+\mathrm{E}(M_k|A=0)\,\mathrm{E}(M_l|A=0)\} \\
&=\{\Sigma_{kl}(1)-\Sigma_{kl}(0)\}+\mathrm{E}(M_k|A=1)\,\mathrm{E}(M_l|A=1)-\mathrm{E}(M_k|A=0)\,\mathrm{E}(M_l|A=0) \\
&\quad+\mathrm{E}(M_k|A=1)\,\mathrm{E}(M_l|A=0)-\mathrm{E}(M_k|A=1)\,\mathrm{E}(M_l|A=0) \\
&=\{\Sigma_{kl}(1)-\Sigma_{kl}(0)\}+\mathrm{E}(M_k|A=1)\{\mathrm{E}(M_l|A=1)\}-\mathrm{E}(M_l|A=0)\} \\
&\quad+\{\mathrm{E}(M_k|A=1)-\mathrm{E}(M_k|A=0)\}\,\mathrm{E}(M_l|A=0) \\
&=\{\Sigma_{kl}(1)-\Sigma_{kl}(0)\}+\mathrm{E}(M_k|A=1)\delta_l+\mathrm{E}(M_l|A=0)\delta_k.
\end{aligned}
$$

Lastly, the direct effect that avoids all t mediators is identified by:

$$
\begin{aligned}
&\mathrm{E}\Bigg[\sum_{m_1,\ldots,m_t}\{\mathrm{E}(Y|1,m_1,\ldots,m_t,C)-\mathrm{E}(Y|0,m_1,\ldots,m_t,C)\}\Pr(M_1=m_1,\ldots,M_t=m_t|A=0,C)\Bigg] \\
&=\beta_A+\sum_{s=1}^{t}\beta_{As}\,\mathrm{E}(M_s|A=0)+\sum_{\substack{k,l=1,\\ k<l}}^{t}\beta_{Akl}\,\mathrm{E}(M_kM_l|A=0).
\end{aligned}
$$