Supplementary Material for

# Sensitivity Analyses for Parametric Causal Mediation Effect Estimation

Jeffrey M. Albert and Wei Wang

Contents: Appendices A-G

## Appendix A. Identifiability of natural direct and indirect effects under mediator comparability (2.3)

Here we show identifiability of the natural direct and indirect effects (1.1) assuming consistency, randomization of $X$ (2.1a), and mediator comparability (2.3), where the latter can be written as

$$E\left\{Y\left(x',m\right) \mid M(x)=m, X=x, L=l\right\} = E\left\{Y\left(x',m\right) \mid M(x')=m, X=x', L=l\right\}. \qquad \text{(A1)}$$

We start by deriving an estimable expression for the conditional expected potential outcome, $E\{Y(x', M(x))| L=l\}$. Noting that $M(x)$ is a random variable, we have, for all $x$, $x'$, $l$ in their respective support sets,

$$E\{Y(x',M(x)) \mid L=l\} = \int_m E(Y(x',m) \mid M(x)=m, L=l) dF_{M(x)|L=l}(m)$$

$$= \int_m E(Y(x',m) \mid M(x)=m, X=x, L=l) dF_{M(x)|L=l}(m) \qquad \text{by (2.1a)}$$

$$= \int_m E(Y(x',m) \mid M(x')=m, X=x', L=l) dF_{M(x)|L=l}(m) \qquad \text{by (A1)}$$

$$= \int_m E(Y \mid M=m, X=x', L=l) dF_{M(x)|L=l}(m) \qquad \text{by consistency}$$

$$= \int_m E(Y \mid M=m, X=x', L=l) dF_{M(x)|X=x,L=l}(m) \qquad \text{by (2.1a)}$$

$$= \int_m E(Y \mid M = m, X = x', L = l) dF_{M|X=x,L=l}(m) \quad . \qquad \text{by consistency}$$

The last expression involves only association model parameters, and thus is identifiable given an appropriate association model (such as (2.4a,b)). The unconditional expected potential outcome $E\{Y(x',M(x))\}$ can then be obtained by integration or summation over $L$. Identifiability of the natural direct and indirect effects (1.1) follows from identifiability of $E\{Y(x',M(x))\}$ for all $x$, $x'$.

Appendix B. Demonstration of identifiability of mediation effects under regression/copula model: Continuous $Y$

For continuous $Y$, we assume the regression relationship (similar to (3.1))

$$E\left\{Y\left(x',m\right) \mid M(x) = m, L = l\right\} = E\left\{Y\left(x',m\right) \mid L = l\right\} + \frac{\sigma_{Y(x',m)}}{\sigma_{M(x)}} \rho_{x'x} \left[m - E\{M(x) \mid L = l\}\right] \quad \text{(B1)}$$

for each $x'$, $x$, $m$ where $\rho_{x'x}$ denotes the correlation between $Y(x',m)$ and $M(x)$, $\sigma^2_{M(x)} \equiv V(M(x) \mid L)$ and $\sigma^2_{Y(x',m)} \equiv V(Y(x',m) \mid L)$; see Section 3 for other notation. We leave $X$ out of the conditioning set in (B1) as this may be dropped due to the assumption of randomized $X$ (2.1a). We assume that $\sigma^2_{Y(x',m)}$ is constant over $m$ and write $\sigma^2_{Y_{x'}} \equiv \sigma^2_{Y(x',m)}$. For continuous (particularly, normally distributed) $Y$ and $M$ we will assume that the variances are homogeneous over $L$. The above regression relationship may be assumed even if $M$ is not continuous (or normally distributed); in this case, we may wish to allow $V(M(x)|L)$ to vary with $L$ (thus, by individual $i$, as discussed below) and use the notation, $\sigma^2_{M(x)i}$.

The difficulty with using (B1) is that the marginal expectation term, $E\{Y(x',m) \mid L=l\}$ (equal to $E\{Y(x',m) \mid X=x', L=l\}$ under randomization ox $X$), is not directly estimable; rather we can estimate $E\{Y(x',m)\mid M(x')=m, X=x', L=l\}$ which takes the form of our association model under the consistency assumption. Note that Wang et al. used formula (B1) directly by essentially assuming, at least in special cases, that $E\{Y(x',m) \mid X=x', L=l\} = E\{Y(x',m)\mid M(x')=m, X=x', L=l\}$, thus allowing the former term to be estimated. In the present approach we avoid any such assumption (which is part of the sequential ignorability assumption). Rather than use (B1) directly we derive an alternative formula by applying (B1) twice (for the cases of $(x', x) = (0,1)$ and $(x', x) = (0,0)$), providing, after some algebraic manipulation,

$$E\left\{Y\left(0,m\right) \mid M(1)=m, L=l\right\} = E\left\{Y\left(0,m\right)\mid M(0)=m, L=l\right\} + \sigma_{Y_0}\left[\frac{\rho\{m-E(M(1)\mid L=l)\}}{\sigma_{M(1)}} - \frac{\rho_0\{m-E(M(0)\mid L=l)\}}{\sigma_{M(0)}}\right]. \tag{B2}$$

The quantity $E\{Y(0, M(1))|L=l\}$ is obtained by integrating the expression in (B2) (over $m$) with respect to the (conditional) probability function of $M(1)$ conditional on $L= l$, denoted as $f\{M(1) \mid L=l\}$. Thus, we have

$$E\left\{Y\left(0,M(1)\right)\mid L=l\right\} = \int E\left\{Y\left(0,m\right)\mid M(1)=m, L=l\right\} f\left(M(1)=m \mid L=l\right\}dm$$

$$= \left[\int E\left\{Y\left(0,m\right)\mid M(0)=m, L=l\right\} f\{M(1)=m \mid L=l\}dm\right] - \sigma_{Y_0}\left[\frac{\rho_0 E\{M(1)-M(0)\mid L=l\}}{\sigma_{M(0)}}\right]. \tag{B3}$$

We then obtain the natural direct effect, $D(1) \equiv E\left\{Y\left(1,M(1)\right)\right\} - E\left\{Y\left(0,M(1)\right)\right\}$, as

$$D(1) = \frac{1}{N_R}\sum\nolimits_{i\in\Pi_R} D(1\mid L=l_i) \tag{B4}$$

where $\Pi_R$ is the reference group of size $N_R$ and, from (B3),

$$D(1\mid L=l_i) \equiv E\left\{Y\left(1,M(1)\right)\mid L=l_i\right\} - E\left\{Y\left(0,M(1)\right)\mid L=l_i\right\}.$$

Identifiability of $D(1)$ (which follows from identifiability of $D(1|L=l_i)$ for each $i$) is obtained by noting that,

$$E\left\{Y\left(1,M(1)\right)\mid L=l_i\right\}-E\left\{Y\left(0,M(1)\right)\mid L=l_i\right\}$$

$$=\int E\left\{Y\left(1,m\right)\mid M(1)=m,L=l_i\right\}f(M(1)=m\mid L=l_i)dm$$

$$-\int E\left\{Y\left(0,m\right)\mid M(0)=m,L=l_i\right\}f(M(1)=m\mid L=l_i\}dm+\eta_0\{E(M(1)-M(0)\mid L=l_i)\}$$

$$=\int E\left\{Y\left(1,m\right)\mid M(1)=m,X=1,L=l_i\right\}f(M(1)=m\mid X=1,L=l_i)dm$$

$$-\left[\int E\left\{Y\left(0,m\right)\mid M(0)=m,X=0,L=l\right\}f\{M(1)=m\mid X=1,L=l\}dm\right]$$

$$+\eta_0[E\{M(1)\mid X=1,L=l_i\}-E\{M(0)\mid X=0,L=l_i\}]$$

$$=\int E\left\{Y\mid M=m,X=1,L=l_i\right\}f(M=m\mid X=1,L=l_i)dm$$

$$-\left[\int E\left\{Y\mid M=m,X=0,L=l\right\}f\{M=m\mid X=1,L=l\}dm\right]+\eta_0[E\{M\mid X=1,L=l_i\}-E\{M\mid X=0,L=l_i\}]$$

$$=E_{M|X=1,L=l_i}E\left\{Y\mid M,X=1,L=l_i\right\}-E_{M|X=1L=l_i}E\left\{Y\mid M,X=0,L=l_i\right\}$$

$$+\eta_0\{E(M\mid X=1,L=l_i)-E(M\mid X=0,L=l_i)\}$$

where $\eta_0=\rho_0\left(\sigma_{Y_0}/\sigma_{M(0)}\right)$, the second equality follows from randomization of $X$, and the third equality follows from consistency; the last equality merely provides a shorthand notation from the definition of (conditional) expectation.

By a similar derivation, the natural indirect effect, $I(0)\equiv E\left\{Y\left(0,M(1)\right)\right\}-E\left\{Y\left(0,M(0)\right)\right\}$, is expressed as,

$$I(0)=\frac{1}{N_R}\sum_{i\in\Pi_R}I(0\mid L=l_i) \qquad \text{(B5)}$$

where

$$I(0 \mid L = l_i) \equiv E\left\{Y\left(0, M(1)\right) \mid L = l_i\right\} - E\left\{Y\left(0, M(0)\right) \mid L = l_i\right\}$$

$$= E_{M|X=1,L=l_i} E\left\{Y \mid M, X = 0, L = l_i\right\} - E_{M|X=0,L=l_i} E\left\{Y \mid M, X = 0, L = l_i\right\}$$

$$-\eta_0\{E(M \mid X = 1, L = l_i) - E(M \mid X = 0, L = l_i)\} .$$

From the above, we can see that under randomization of $X$ and specified association models for $Y$ and $M$ (as in (2.4a,b)), $D(1)$ and $I(0)$ are identifiable once $\eta_0$ is given. Thus, a sensitivity analysis in this linear case may be obtained by varying $\eta_0$. Alternatively, and perhaps more usefully, we can estimate $\sigma^2_{Y_0}$ and $\sigma^2_{M(0)}$ under homogeneous variance assumptions, and use $\rho_0$ as the sensitivity parameter. Specifically, one approach would be to assume, for example, that $\sigma^2_{Y_0} = \sigma^2_{Y_0|M(0)} / (1 - \rho_0{}^2)$, where $\sigma^2_{Y_0|M(0)} \equiv V\{Y(0,m) \mid M(0), L = l\}$, assumed to be homogenous over $m$, is estimable.

The simple expressions for $D(1)$ and $I(0)$ provide in Section 3.1 are obtained in the case where $Y$ and $M$ follow a linear additive regression models, that is,

$$E(Y \mid X, L, M) = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 L , \qquad E(M \mid X, L) = \gamma_0 + \gamma_1 X + \gamma_2 L . \tag{B6}$$

In this case we have

$$E_{M|X=1,L=l_i} E(Y \mid M, X = 1, L = l_i) = \beta_0 + \beta_1 + \beta_2 E(M \mid X = 1, L = l_i) + \beta_3 l_i$$

and

$$E_{M|X=1L=l_i} E(Y \mid M, X = 0, L = l_i) = \beta_0 + \beta_2 E\{M \mid X = 1, L = l_i\} + \beta_3 l_i$$

so that, under randomization of $X$ (2.1a), we have

$$D(1 \mid L = l_i) = \beta_0 + \beta_1 + \beta_2 E\{M \mid X = 1, L = l_i\} + \beta_3 l - [\beta_0 + \beta_2 E\{M \mid X = 1, L = l_i\} + \beta_3 l]$$
$$+ \eta_0\{E(M \mid X = 1, L = l_i) - E(M \mid X = 0, L = l_i)\}$$

$$= \beta_1 + \eta_0 \gamma_1 = D(1) . \tag{B7}$$

Similarly,

$$I(0 \mid L = l_i) = \beta_0 + \beta_2 E\{M \mid X = 1, L = l_i\} + \beta_3 l_i - [\beta_0 + \beta_2 E\{M \mid X = 0, L = l_i\} + \beta_3 l_i]$$
$$- \eta_0 \{E(M \mid X = 1, L = l_i) - E(M \mid X = 0, L = l_i)\}$$

$$= (\beta_2 - \eta_0)\gamma_1 = I(0). \qquad \text{(B8)}$$

Note that when *M* is discrete, or when one might otherwise wish to use a generalized linear model for *M* with a non-identity link, we may use the following more general expressions, allowing expected values and variances of *M* to depend on *L* (and *X* if desired):

$$D(1) = \beta_1 + \sigma_{Y_0} \rho_0 (1/N_R) \sum_{i \in \Pi_R} (\gamma_{1i} / \sigma_{M(0)i})$$

$$I(0) = (1/N_R) \left\{ \beta_2 \sum_{i \in \Pi_R} \gamma_{1i} - \sigma_{Y_0} \rho_0 \sum_{i \in \Pi_R} (\gamma_{1i} / \sigma_{M(0)i}) \right\}$$

where $\gamma_{1i} \equiv E(M \mid X = 1, L = l_i) - E(M \mid X = 0, L = l_i)$. The latter expression may be estimated based on the assumed model for *M*.

## Appendix C. Demonstration of identifiability of mediation effects under regression/copula model: Discrete *Y*

To begin, we suppose that the observed *M* is continuous, in which case we use $M^*(x) = M(x)$ and define $m_{sx} \equiv [m - E\{M(x)\}] / \sigma_{M(x)}$ for each *x*. We assume the regression relationship (3.1) and thus (3.2) for observed *M*, that is, $E\left\{Y^*\left(x', m\right) \mid M(x) = m\right\} = \rho_{x'x} m_{sx}$. From our assumption of bivariate normality of $Y^*(x', m)$ and $M(x)$ (with correlation $\rho_{x'x}$ and standard normal marginal distribution of $Y^*(x', m)$), it follows that $Y^*(x', m) \mid \{M(x) = m, X = x, L = l\} \sim \mathrm{N}(\rho_{x'x}\, m_{sx}, 1 - \rho_{x'x}^2)$. We let $p_{0ji} \equiv P\{Y(0,m) = j \mid M(0) = m, X = 0, L = l_i\}$ though later drop the individual indicator, *i*, as well

as the conditioning on $X$, from the notation. Our goal is to identify and estimate the probabilities, $p_j \equiv P\{Y(0,m) = j \mid M(1)= m, X=0, L=l\}$, $j=1,\ldots,J$, which are needed to estimate $E\{Y(0,M(1))\}$. We denote cumulative probabilities as $P_{0j} \equiv P\{Y(0,m) \le j \mid M(0) = m, X=0, L=l\}$ and $P_j \equiv P\{Y(0,m) \le j \mid M(1) = m, X=0, L=l\}$.

First, we connect the latent variable $Y^*$ to the observed $Y$ through the following correspondence:

$$Y(0,m) = j \quad \Leftrightarrow \quad Y^*_{0,j-1} < Y^*(0,m) \le Y^*_{0j}$$

for $j = 1,\ldots,J$, where $Y^*_{0j} = (1-\rho_0{}^2)^{\frac{1}{2}}\Phi^{-1}(P_{0j}) + \rho_0 m_{s0}$, $Y^*_{0,-1} \equiv -\infty$ and $Y^*_{0J} = \infty$. It is easy to check, by transforming $Y^*(0,m)| \{M(0)=m, X=0, L=l\} \sim N(\rho_0\, m_{s0}, 1\text{-}\rho_0{}^2)$ to a standard normal variate, that $P\{Y^*(0,m) \le Y^*_{0j} \mid M(0) = m, X = 0, L = l\} = = \Phi\left\{\dfrac{Y^*_{0j} - \rho_0 m_{s0}}{(1-\rho_0{}^2)^{\frac{1}{2}}}\right\} = P_{0j}$.

Then, since $Y^*(0,m)|M(1)=m \sim N(\rho\, m_{s1}, 1\text{-}\rho^2)$, we have

$$\begin{aligned}
P_j &\equiv P\left\{Y(0,m) \le j \mid M(1) = m, X = 0, L = l\right\} \\
&= P\left\{Y^*(0,m) \le Y^*_{0j} \;\middle|\; M(1) = m, X = 0, L = l\right\} \\
&= P\left\{\frac{Y^*(0,m) - \rho m_{s1}}{(1-\rho^2)^{\frac{1}{2}}} \le \frac{Y^*_{0j} - \rho m_{s1}}{(1-\rho^2)^{\frac{1}{2}}} \,\middle|\, M(1) = m, X = 0, L = l\right\} \\
&= \Phi\left\{\frac{Y^*_{0j} - \rho m_{s1}}{(1-\rho^2)^{\frac{1}{2}}}\right\} \\
&= \Phi\left\{\frac{(1-\rho_0{}^2)^{\frac{1}{2}}\Phi^{-1}(P_{0j}) + \rho_0 m_{s0} - \rho m_{s1}}{(1-\rho^2)^{\frac{1}{2}}}\right\}. \qquad \text{(C1)}
\end{aligned}$$

From (C1) we see that $\rho = \rho_0 = 0$ yields $P_{0j} = P_j$ for each $j$. Under randomization of $X$ (2.1a) this result in turn implies mediator comparability (2.3) for $x' = 0$ and thus identifiability of the natural direct and indirect effects, $D(1)$ and $I(0)$.

Given the $P_j$'s for each individual from (C1), written as $P_{ji}$, we can write the expected value for an individual ($i$) as,

$$E\{Y(0,m) \mid M(1) = m, X = 0, L = l_i\} = E\{Y(0,m) \mid M(1) = m, L = l_i\} = \sum_{j=1}^{J} j(P_{ji} - P_{j-1,i}) \qquad \text{(C2)}$$

where the first equality follows from our assumption of randomization of $X$ (2.1a). Note that under the copula model leading to (C1), with an accompanying association model (such as (2.4a,b)), the $P_j$'s are estimable for each individual (that is, each value of $l_i$) for a given value for $m$ (along with $\rho$ and $\rho_0$). To estimate the marginal expected value for an individual, i.e., $E\{Y(0,M(1)) \mid L_i = l_i\}$, we take the expectation of (C2) with respect to $m$ (that is, the distribution of $M(1)$). As (C2) is nonlinear in $m$ this would need to be done by integration. Alternatively, we can use a Monte Carlo approach (as outlined in Section 3) by computing the average of (C2) plugging in independent replicates of $m$ generated from its assumed distribution, namely, normal, with mean and variance obtained from model (2.4b), for each individual (or $m_{s1}$ generated from a standard normal distribution). Finally, to obtain the population estimate of $E\{Y(0,M(1))\}$ we take the average of the individual estimates in the selected reference group.

For discrete $M$, the above algorithm may still be used if one is willing to assume that the regression relationship (3.1) holds with the observed $M$. Note that in this case (C2) can be used as before and $E\{Y(0,M(1))\}$ obtained by summing over the discrete distribution for $M(1)$ (based on its model) for each person, and over the empirical distribution for $L$, giving, for example,

$$E\{Y(0,M(1))\} = \sum_{i=1}^{N_R} \sum_{k=1}^{K} \left\{ P(M = k \mid X = 1, L = l_i) \sum_{j=1}^{J} j(P_{ji} - P_{j-1,i}) \right\}$$

substituting $k$ for $m$ in the formula for $P_{ji}$.

Otherwise, to use a latent variable ($M^*$) for a discrete $M$ (with values $k = 0,\ldots,K$) the above algorithm needs modification. In this case, since $Y^*(x',m) \mid \{M^*(x)= m^*, X=x, L=l\} \sim N(\rho_{x'x}\, m_{sx}^*, 1-\rho_{x'x}^2)$, we have, for appropriately chosen individual- (and $k$-) specific cutpoints, $Y_{0kj}^*$, for $j = 0,\ldots, J-1$,

$$P_{0kj} \equiv P\left\{Y(0,k) \le j \mid M(0)=k, X=0, L=l\right\}$$

$$= P\left\{Y^*(0,k) \le Y_{0kj}^* \;\middle|\; m_{0,k-1}^* \le M^*(0) = m_0^{\,*} < m_{0,k}^*, X=0, L=l\right\}$$

$$= \frac{\int_{m_0^* \in \Omega_{0k}} P\{Y^*(0,k) \le Y_{0kj}^* \mid M^*(0)=m_0^{\,*}, X=0, L=l\} dF_{M^*(0)=m_0^{\,*}\mid X=0, L=l_i}(m_0^{\,*})}{\int_{m_0^* \in \Omega_{0k}} dF_{M^*(0)=m_0^{\,*}\mid X=0, L=l_i}(m_0^{\,*})}$$

$$= \frac{\int_{m_0^* \in \Omega_{0k}} \Phi\left\{\dfrac{Y_{0kj}^* - \rho_0 m_{s0}^*}{(1-\rho_0^{\,2})^{\frac{1}{2}}}\right\} dF_{M^*(0)=m_0^{\,*}\mid X=0, L=l_i}(m_0^{\,*})}{\int_{m_0^* \in \Omega_{0k}} dF_{M^*(0)=m_0^{\,*}\mid X=0, L=l_i}(m_0^{\,*})} \qquad \text{(C3)}$$

where $\Omega_{0k} = (m_{0,k-1}^*, m_{0k}^*]$, $m_{0k}^* = \Phi^{-1}\{P(M(0) \le k \mid X=0, L=l)\}$ for $k$=0,…,$K$-1, $m_{0,-1}^* = -\infty$, and $m_{0K}^* = \infty$. The cutpoints, $Y_{0kj}^*$ ($j$= 0,…$J$-1), for each $k$ are obtained so that (C3) holds, noting that the $P_{0kj}$ and $m_{0k}^*$ are estimable. This may be solved numerically, reasonable starting values being $Y_{0kj}^* = (1-\rho_0^{\,2})^{\frac{1}{2}}\Phi^{-1}(P_{0kj}) + \rho_0 m_{0kA}^*$, for $j$= 0,…$J$-1, where $m_{0kA}^* = \Phi^{-1}[\{\Phi(m_{0k}^*) + \Phi(m_{0,k-1}^*)\}/2]$, $Y_{0k,-1}^* \equiv -\infty$ and $Y_{0kJ}^* = \infty$.

Then, we have

$$P_{kj} \equiv P\{Y(0,k) \le j \mid M(1)=k, X=0, L=l\}$$

$$= P\left\{Y^*(0,k) \le Y_{0kj}^* \;\middle|\; m_{1,k-1}^* \le M^*(1) = m^* < m_{1,k}^*, X=0, L=l\right\}$$

$$= \frac{\int_{m^* \in \Omega_{1k}} P\{Y^*(0,k) \le Y_{0kj}^* \mid M^*(1)=m^*, X=0, L=l\} dF_{M^*(1)=m^*\mid X=0, L=l}(m^*)}{\int_{m^* \in \Omega_{1k}} dF_{M^*(1)=m^*\mid X=0, L=l}(m^*)} \qquad \text{(C4)}$$

where $\Omega_{1k} = (m^*_{1,k-1}, m^*_{1,k}]$ and $m^*_{1,k} = \Phi^{-1}\{P(M(1) \le k \mid X = 0, L = l)\}$, the latter being estimable under randomization of $X$ (2.1a) and consistency. Note that the denominator in (C4), which we denote as $P_{Mk}$, is equal to $P\{M(1)=k \mid X=0, L=l\} = P\{M=k \mid X=1, L=l\}$ (under randomization of $X$ and consistency), the latter term being estimable from model (2.4b). Since we assume $Y^*(0,m)|\{M^*(1)=m^*, X=0, L=l\} \sim N(\rho\, m_{s1}{}^*, 1-\rho^2)$, a similar approach to above (leading to (C1)) yields

$$P\left\{Y^*(0,k) \le Y^*_{0kj} \;\middle|\; M^*(1) = m^*, X = 0, L = l\right\}$$

$$= \Phi\left\{\frac{Y^*_{0kj} - \rho m^*_{s1}}{(1-\rho^2)^{\frac{1}{2}}}\right\} \equiv P^*_{kj} \qquad \text{(C5)}$$

keeping in mind that $P^*_{kj}$ is a function of $m^*$. Substituting the expression (C5) for the corresponding probability in (C4) gives an estimable expression for $P_{kj}$ for given $\rho_0$, $\rho$ for each individual, namely,

$$P_{kj} = \frac{1}{P_{Mk}} \int_{m^* \in \Omega_{1k}} P^*_{kj}\, dF_{M^*(1)=m^* \mid X=0, L=l_i}(m^*). \qquad \text{(C6)}$$

To compute (C6), we would need to either integrate over the specified (conditional) distribution of $M^*(1)$ or use a Monte Carlo approach. In a Monte Carlo approach to computing (C4), we draw, for each individual, $R$ independent replicates of $M^*(1)$, with the $r$th value denoted by $m_{ir}{}^*$, from $N(0, 1)$. We obtain an estimate of $P_{kj}$ for individual $i$ by computing (with hats indicating model-based estimates, and the indices $i$ and $r$ added for individual and replicate, respectively),

$$\hat{P}_{kji} = \frac{1}{N_{\hat{\Omega}_{1k}}} \sum_{m^*_{ir} \in \hat{\Omega}_{1k}} \hat{P}^*_{kjir}$$

where $N_{\hat{\Omega}_{1k}}$ is the number of $m_{ir}{}^*$ that fall into the interval $\hat{\Omega}_{1k}$. Next, we use an expression analogous to (C2), to get $\hat{E}\{Y(0,k) \mid M(1) = k, L_i = l_i\} = \sum_{j=1}^{J} j(\hat{P}_{kji} - \hat{P}_{k,j-1,i})$.

We then compute

$$\hat{E}\{Y(0,M(1)) \mid L_i = l_i\} = \sum_{k=1}^{K} \hat{E}\{Y(0,k) \mid M(1)=k, L=l_i\}\hat{P}\{M(1)=k \mid L=l_i\}$$

$$= \sum_{k=1}^{K} \left\{ \sum_{j=1}^{J} j(\hat{P}_{kji} - \hat{P}_{k,j-1,i}) \right\} \hat{P}_{Mki} \ .$$

Finally, we average over the $l_i$ for the selected reference group ($\Pi_R$) to get

$$\hat{E}\{Y(0,M(1))\} = \sum_{i \in \Pi_R} \sum_{k=1}^{K} \left\{ \sum_{j=1}^{J} j(\hat{P}_{jki} - \hat{P}_{j-1,ki}) \right\} \hat{P}_{Mki}$$

from which estimates of $D(1)$ and $I(0)$ are readily obtained.

Appendix D. Demonstration of identifiability of mediation effects under the hybrid model (3.5)

As noted in Appendix A, identifiability of the natural direct and indirect effects follows from identifiability of the conditional expected potential outcome, $E\{Y(x', M(x)) \mid L=l\}$. Assuming the hybrid model (3.5) along with previous assumptions (but dropping the mediator comparability assumption (2.3)) we have

$$E\{Y(x', M(x)) \mid L=l\} = \int_m E(Y(x', m) \mid M(x)=m, L=l) dF_{M(x)|L=l}(m)$$

$$= \int_m E(Y(x', m) \mid M(x)=m, X=x, L=l) dF_{M(x)|X=x,L=l}(m) \qquad \text{by (2.1a)}$$

$$= \int_m E(Y(x', m) \mid M(x)=m, X=x, L=l) dF_{M|X=x,L=l}(m) \qquad \text{by consistency}$$

$$= \int_m h_2^{-1}\left\{ \beta_0 + \beta_1\left(\phi x + (1-\phi)x'\right) + \beta_2 m + \beta_3 l \right\} dF_{M|X=x,L=l}(m) \qquad \text{by (3.5)}$$

Note that the $\beta$'s in the last expression are estimable from the association model that is a special case of the hybrid model with $x = x'$. The last term (the probability function for $M$ given $X$ and $L$)

is estimable from the model for $M$. We thus see that $\phi$, the sensitivity parameter, is the only non-estimable parameter in the last expression. Upon specification of a value for $\phi$ we thus obtain an estimable expression for, that is, identifiability of $E\{Y(x', M(x))| L=l\}$, consequently $E\{Y(x', M(x)\}$ by summing over the empirical distribution of $L$, and thus the natural direct and indirect effects (1.1).

As an addition note, it is easy to see that the mediator comparability assumption (added to the previous assumptions, including randomization of $X$) implies that $\phi = 0$ in the hybrid model. First, it can be seen, by applying the link function, $h_2$, to both sides of the hybrid model (3.5) and plugging in appropriate values for $x'$ and $x$, that

$$\beta_1\phi = h_2\left[E\left\{Y\left(0,m\right)|M(1)=m,X=1,L=l\right\}\right]-h_2\left[E\left\{Y\left(0,m\right)|M(0)=m,X=0,L=l\right\}\right]$$

so that

$$\phi=\left(h_2\left[E\left\{Y\left(0,m\right)|M(1)=m,X=1,L=l\right\}\right]-h_2\left[E\left\{Y\left(0,m\right)|M(1)=m,X=0,L=l\right\}\right]\right)/\beta_1.$$

From this expression, we see that mediator comparability (2.3), written (with $x' = 0$) as

$$E\left\{Y\left(0,m\right)|M(1)=m,X=1,L=l\right\}=E\left\{Y\left(0,m\right)|M(0)=m,X=0,L=l\right\},$$

implies that $\phi = 0$.

The hybrid model thus provides a general structure that essentially replaces (and generalizes) the mediator comparability (or the second sequential ignorability) assumption. When $\phi = 0$ (equivalent to mediator comparability) the hybrid model essentially reduces to the corresponding (estimable) association model. To see this, note that $\phi = 0$, plugged into the hybrid model, implies that

$$E\{Y(x',m) \mid M(x) = m, X = x, L=l\} = h_2^{-1}\{\beta_0 + \beta_1 x' + \beta_2\, m + \beta_3\, l\}.$$

As the right hand expression does not depend on $x$, we have

$$E\{Y(x',m) \mid M(x) = m, X = x, L=l\} = E\{Y(x',m) \mid M(x') = m, X = x', L=l\}$$

$$= E\{Y \mid M = m, X = x', L=l\}$$

the last equality following from consistency. Thus, when $\phi$=0 in the general context of the hybrid model, the expected potential outcome underlying the natural direct and indirect effects can be equated to an estimable regression function.

## Appendix E. Guidelines for Elicitation of Sensitivity Parameters

In this appendix, we describe a procedure for the specification, or elicitation, of the sensitivity parameters for both the copula model and hybrid model approaches. We start with the later, in part because the nature of the parameters may be less familiar to most users, and because the results may be used for the copula model specification.

### I. Hybrid Model Approach

It may be difficult to specify the hybrid model sensitivity parameter, $\phi$, directly through its interpretation as the proportion of the direct effect parameter $\beta_1$ due to selection bias. (Recall, specifically, that $\beta_1$ is a controlled direct effect, that is, the estimable effect, on the scale of the link function, of exposure ($X$) on the final outcome ($Y$) conditional on the observed mediator and included baseline covariates.) In particular, the magnitude, or even the direction, of the selection bias (also referred to as 'collider-stratification bias'), and its connection to confounding, may not be intuitive to many researchers. Consequently, in this appendix, we present a simple step-by step approach that allows elicitation of $\phi$ using more basic, intuitive, elements involving effects of unobserved confounders.

Our suggested approach, providing an approximation to the parameter in the original scale, involves the use of a linear scale for the elicitation of $\phi$. Note that, as we assume exchangeability of levels of the exposure variable $X$, the main concern (and source of violation of sequential ignorability) is unobserved $M$-$Y$ confounders. Specifically, we consider the linear regression of $Y$ on $X$ , $M$, and $L$ and denote the coefficient of $X$ in this model as $\beta_{Y \cdot X|M}$. In general $\beta_{Y \cdot X|M}$ will not be the same as the true direct effect of $X$ on $Y$ , denoted as $\beta_{Y \cdot X|M,U}$, which would be estimable if the unmeasured $M$-$Y$ confounders (denoted by the vector $U$) were included in the above model. We will refer to this discrepancy as the 'bias', $B^*$; that is, we let $B^* \equiv \beta_{Y \cdot X|M} - \beta_{Y \cdot X|M,U}$. Similarly, we let $\beta_{M \cdot X}$ denote the coefficient of $X$ in the linear regression of $M$ on $X$ and $L$, and let $\beta_{M \cdot X|U}$ denote the direct effect of $X$ on $M$ (i.e., the coefficient of $X$), when $U$ is also included (along with $L$) in the model.

Of course, in general (unless the generalized linear model for $Y$ involves an identity link function), the direct effect, $\beta_{Y \cdot X|M,U}$, from the linear model is not the same as the corresponding parameter (i.e., $\beta_1$) in the generalized linear model of interest. However, for the elicitation of the sensitivity parameter, we suppose that the 'relative bias', $(\beta_{Y \cdot X|M} - \beta_{Y \cdot X|M,U}) / \beta_{Y \cdot X|M}$ , from the linear model, provides a reasonable approximation of the corresponding quantity (namely, $\phi$) for the assumed generalized linear model. We may estimate $\beta_{Y \cdot X|M}$ using the ordinary least squares estimator, $b_{Y \cdot X|M}$ . However, $\beta_{Y \cdot X|M,U}$ , and thus the bias (or relative bias), even on the linear scale, is not estimable. The present goal is to provide a user-friendly approach to elicit plausible values for the bias (and thus $\phi$) based on expert subject-matter knowledge. The proposed steps are as follows:

1. Compute $b_{Y \cdot X|M}$, $b_{M \cdot X}$, the ordinary least squares estimates of $\beta_{Y \cdot X|M}$, $\beta_{M \cdot X}$, respectively.

2. List possible candidate (unobserved, *M*-*Y*) confounders. Label these (ideally, in order of 'strongest' to 'weakest') as $U_1,..,U_K$.

3. For the first candidate confounder, $U_1$, specify (estimated or guessed) values for $b_{Y \cdot U1 \,|M,\, X}$ and $b_{M \cdot U1|\, X}$, the coefficient of $U_1$ in the extended regression model for *Y* and *M*, respectively. Note, as shown in Figure E1 below, that these coefficients are analogous to the estimated regression coefficients for *X* (obtained in Step 1), but now assuming that *X* is already in the model. From an examination of the paths in Figure E1, we write the contribution to the estimate of the bias due to the confounder $U_1$ as,

$$B_1 = - \, b_{M \cdot X} \;\; b_{Y \cdot U1 \,|\, M,\, X} \;\; b_{M \,\cdot U1 \,|\, X}.$$

Note that, as expected for selection bias, a confounder whose effects are in the same direction as those of the exposure would induce a negative bias. Note also that the corresponding bias component for the indirect effect would then be $-B_1$.

For subsequent confounders, coefficients are specified in sequence assuming that the exposure and previous confounders are included in the model. Thus, for the *k*th confounder we specify the analogous quantities, $b_{Y \cdot U\mathrm{k} \,|M,X,U1,..,Uk\text{-}1}$ and $b_{M \,\cdot U\mathrm{k} \,|X,U1,\ldots,Uk\text{-}1}$, and the additional bias due to this confounder is

$$B_k = - \, b_{M \cdot X} \;\; b_{Y \cdot U\mathrm{k} \,|\, M,\, X} \;\; b_{M \cdot U\mathrm{k} \,|\, X}$$

dropping the conditioning on previous *U*'s in the notation for brevity.

4. Compute the total bias as

$$B = \Sigma_k \, B_k = - \, b_{M \cdot X} \, \Sigma_k \, b_{Y \cdot U\mathrm{k} \,|\, M,\, X} \;\; b_{M \cdot U\mathrm{k} \,|\, X}$$

and obtain the specified sensitivity parameter value as $\phi = B \,/\, b_{Y \cdot X \,|\, M}$, representing the (selection) bias as a proportion of the estimable (controlled) direct effect.

The above approach may be refined, to obtain a more accurate specification of the bias, by specifying $\sigma^2_{Uk|M,X} \equiv V(U_k \,|M, X)$ and $\sigma^2_{M|Uk,X} \equiv V(M \,|U_k, X)$ and using $b^c_{M \cdot Uk \,|\, X} \equiv b_{M \cdot Uk \,|\, X}$ $(\sigma^2_{Uk|M,X} / \sigma^2_{M|Uk,X})$ in place of $b_{M \cdot Uk \,|\, X}$ throughout (for $k$=1,…,$K$), where the proper conditioning on previous $U$'s (not in the notation) is done for $k$>1. Note that $b^c_{M \cdot Uk \,|\, X}$ is equal to $b_{Uk \cdot M|\, X}$ (the coefficient of $M$ in the regression of $U_k$ on $M$ and $X$) which is the appropriate term for determining the path effect, but may be less intuitive (and causally meaningful), and thus more difficult to specify directly, than $b_{M \cdot Uk \,|\, X}$.

We note that the above approach accommodates different types of unobserved confounders (for example, binary, ordinal, or continuous), as well as $M$ and $Y$. In practice, we may wish to consider multiple specifications (or scenarios) and thus obtain multiple values, or a range, for $\phi$.

An alternative to specifying parameter values for multiple confounders is to consider a single composite confounder. Using a composite confounder would allow one to follow the elicitation procedure described above for the case of a single confounder. This approach has the advantages of simplicity and of allowing one to avoid having to specify the effects of confounders conditional on proceeding confounders. On the other hand, some users may have difficulty in conceiving of and specifying the effect of a composite variable. Of course, both approaches may be tried and compared to arrive at a satisfactory specification.

To provide a numerical example, we detail the elicitation of the sensitivity parameter value for the dental data analysis (as given in Section 4), conducted with the assistance of our dental colleague, Dr. Suchitra Nelson.

1. From the fit of the linear model of DMFTD ($Y$) regressed on MomEd ($X$), OHI ($M$) and baseline covariates ($L$) we obtained the estimate, $b_{Y \cdot X \,|\, M} = 0.12$, and from the regression of OHI on MomEd and $L$ we obtained $b_{M \cdot X} = 0.29$. See the corresponding causal diagram in

Figure E1 keeping in mind that the filled in value for the direct effect of $X$ on $Y$ ($b_{Y \cdot X \mid M}$ = 0.12) may be a biased estimate.

2. Potential $M$-$Y$ confounders (from 'strongest' to 'weakest') are financial means (1 for low, 0 for high), stress (1 for high, 0 for low), and social support (1 for low, 0 for high).
3. From prior substantive knowledge, we specify the regression coefficients for the three potential confounders as follows:

| $k$ | confounder | $b_{Y \cdot Uk \mid M, X}$ | $b_{M \cdot Uk \mid X}$ |
|---|---|---|---|
| 1 | financial | 0.2 | 0.3 |
| 2 | stress | 0.1 | 0.1 |
| 3 | social supp | 0.1 | 0.1 |

Note that these specified values replace the question marks in Figure E1.

4. We compute the bias as $B$ = -0.29{(0.2)(0.3) + (0.1)(0.1) + (0.1)(0.1)} ≈ -0.023, resulting in a specified sensitivity parameter value of $\phi = B / b_{Y \cdot X \mid M}$ = -0.023/ 0.12 ≈ -0.2.

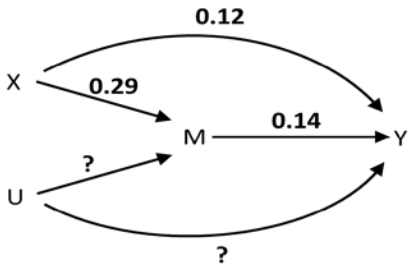


Figure E1 – DAG and Regression Estimates for Dental Data

To use the refinement mentioned above, we specify, for example, ($\sigma^2_{U1|M,X} / \sigma^2_{M|U1,X}$) = 0.92 and thus $b^c_{M \cdot U1 \mid X} \equiv b_{M \cdot U1 \mid X}$ (0.92) = 0.3 (0.92) = 0.28. This specification was obtained from simulated data mimicking the dental data and using a binary unobserved confounder ($U$) with $P(U=1)$ = 0.5. Similar corrections may be used for the other unobserved confounders. Because

this correction is not substantial (particularly relative to the uncertainty in the specification of the regression coefficients) in the present case, we do not use it in the specifications for this example.

As an alternatively to the above specification, we may wish to consider a more pessimistic scenario using coefficient values ($b_{Y \cdot Uk \mid M, X}$ and $b_{M \cdot Uk \mid X}$) for 'financial means' of 0.3 and 0.4, and for 'stress' of 0.2 and 0.15. This scenario would yield $B$= - 0.29{(0.3)(0.4) + (0.2)(0.15) + (0.1)(0.1)} = -0.046 and $\phi$ = -0.046 / 0.12 ≈ -0.4. For additional conservatism, we suppose a 2.5-fold-increase in $\phi$ from our more pessimistic scenario, resulting in a value of around $\phi$ = -1. Thus, we consider a plausible range for $\phi$ to be -1 to 0 (including as an upper bound the most optimistic scenario of no *M*-*Y* confounding).

This approach supposes that the relative bias from the linear model transports reasonably well to a generalized linear model with a non-identity link function. In many applications it will be not essential that this approximation be very accurate, as only a rough idea of plausible values for the sensitivity parameter is needed. However, some simulation results provided in Supplementary Material Appendix F suggest that this approximation may be quite good.

An alternative approach that would be more accurate in terms of the connection between the specified parameter values and the actual $\phi$, would be to use the link function scale (e.g., logit for logistic regression) for the regression coefficients to be specified/elicited. The elicitation would follow the above approach except on the alternative scale.

The above approach may resemble previous sensitivity analysis approaches in its consideration of unobserved confounders. However, there are some important distinctions and advantages of the hybrid model approach in conjunction with the above method for sensitivity parameter elicitation, namely, 1) it can accumulate the impact of multiple hypothesized (*M*-*Y*) confounders, and 2) it allows a choice of scales for the specification of confounder effects.

II. Copula Model Approach

The sensitivity parameters for copula model are particular instances of the correlation $\rho_{x',x} \equiv \text{corr}\{Y^*(x',m), M^*(x)\}$. For estimation of $D(1)$ and $I(0)$, as shown in Appendices B and C, the particular sensitivity parameters are $\rho \equiv \rho_{01} \equiv \text{corr}\{Y^*(0,m), M^*(1))\}$ and $\rho_0 \equiv \rho_{00} \equiv \text{corr}\{Y^*(0,m), M^*(0)\}$, or for the additive linear $Y$ model, $\rho_0$ alone.

To help provide intuition for the specification of values for $\rho$ and $\rho_0$ we note that these correlations (conditional on baseline covariates, $L$) represent an effect of $M$-$Y$ confounding (or selection bias). For example, $\text{corr}\{Y^*(0,m), M^*(0)\}$ is a correlation of $Y^*(0,m)$ with $M^*(0)$ which is not due to the causal effect of $M$ since $Y^*(0,m)$ is the (potential) outcome due to setting $M = m$ (and $X = 0$); note that the levels of $M^*(0)$ correspond to subgroups (or cohorts) of the population. When unobserved confounders are positively related to $M$ and $Y$ (as is the case for low financial status, which is positivity related both to OHI and DMFTD), then $\text{corr}\{Y^*(0,m), M^*(0)\}$ will be positive. If there are interactions between such confounders and $X$, then the correlations $\rho_{x',x}$ may vary according to $x'$ and $x$ (though the correlations may be heterogeneous for other reasons as indicated below).

Although such correlations (which are, of course, bounded by -1 and 1) can be specified directly, the quantitative specification may be assisted by considering, in a similar manner as for the hybrid model, the connection between the copula model correlations and the effects of potential confounders. One relatively simple approach is along the lines of an approach presented, in the linear model context, by Imai et al., (2010b).

In this approach we suppose the following set of linear structural models (where, despite the common notation used for convenience, there is no relationship between the regression parameters involved here and those given previously):

$$Y^*(x', m) = \beta_0 + \beta_1 x' + \beta_2 m + \beta_3 l + \varepsilon_Y{}^*(x', m)$$

$$M^*(x) = \gamma_0 + \gamma_1 x + \gamma_2 l + \varepsilon_M{}^*(x) .$$

The conditional correlation, $\rho_{x'x}$, of $Y^*(x',m)$ and $M^*(x)$ under these models is the same as corr$\{\varepsilon_Y^*(\underline{x'},m), \varepsilon_M^*(x)\}$. Extending Imai et al. (2010b), we suppose the decompositions,

$$\varepsilon_Y{}^*(x', m) = \lambda_{Yx'} U + \varepsilon_Y \quad \text{and} \quad \varepsilon_M{}^*(x) = \lambda_{Mx} U + \varepsilon_Y$$

where $\lambda_{Yx'}$, $\lambda_{Mx}$ are fixed unknown coefficients, $U$ is random (representing an unobserved confounder) and $\varepsilon_Y$, $\varepsilon_M$, and $U$ are independent.

Let $R_{Y(x')}{}^2$, $R_{M(x)}{}^2$ be the coefficient of determination (for $U$) in the regression of $\varepsilon_Y^*(\underline{x'},m)$ and $\varepsilon_M^*(x)$, respectively. Note that we assume that $R_{Y(x')}{}^2$ does not depend on the mediator value, $m$; thus the latter is not included in the notation. A given $R^2$ has an appealing interpretation as the proportion of variance (of $Y^*$ or $M^*$) explained by the confounder. Thus, a relatively easy way to elicit a value for $\rho_{x'x}$ (for given $x'$ and $x$ of interest) may be to specify the relevant $R^2$'s (namely, $R_{Y(x')}{}^2$ and $R_{M(x)}{}^2$) and use the following relationship (Imai et al., 2010b),

$$\rho_{x'x} = \text{sgn}(\lambda_{Yx'} \cdot \lambda_{Mx}) R_{Y(x')} \cdot R_{M(x)}. \qquad \text{(E1)}$$

We illustrate this approach by carrying out the elicitation of the copula model correlations for our dental data example. First we consider possible *M*-*Y* confounders. This was done above in our elicitation based on the hybrid model where we indicated that the strongest unobserved confounder was thought to be 'financial means'. As the present approach involves a single confounder, we can consider financial means alone (conceived as either binary or continuous) or imagine a summary measure or latent variable encapsulating all the candidate confounders.

For the estimands $D(1)$ and $I(0)$, we need to elicit values for $\rho_{00}$ and $\rho_{01}$. Our specifications (for a 'pessimistic' scenario as well as a plausible range) of the $R^2$s for 'financial means' (with each given outcome) are given in the following table:

| Outcome | Outcome Description | $R^2$ (pessimis) | $R^2$ range |
|---|---|---|---|
| $Y(0,m)$ | DMFT for high MomEd | 0.5 | (0.1, 0.7) |
| $Y(1,m)$ | DMFT for low MomEd | 0.55 | (0.1, 0.75) |
| $M(0)$ | OHI for high MomEd | 0.5 | (0.1, 0.7) |
| $M(1)$ | OHI for low MomEd | 0.6 | (0.1, 0.8) |

The $R^2$ for $Y(1,m)$ is not actually needed for the present estimands (rather, would be needed for the alternative estimands, $D(0)$ and $I(1)$), but is included for reference.

We note that the signs of the regression coefficients $\lambda_{Y0}$, $\lambda_{M0}$, and $\lambda_{M1}$ are all expected to be positive for financial means (defined such that a higher score is worse - 'less' financial means). Also note that although the $Y^*$ and $M^*$ represent underlying, generally unobserved, continuous variables, considering the observed outcomes may provide a reasonable approximation. In the case of the present example, $M$ (OHI) is roughly continuous anyway (and the analysis uses $M^* = M$) and $Y$ (DMFTD) is a dichotomization of an underlying count variable (which may itself be considered, and is perhaps not far from continuous).

We then have, using the 'pessimistic' values in the above table, $\rho_{00} = \text{sgn}(\lambda_{Y0} \cdot \lambda_{M0}) R_{Y(0)} \cdot R_{M(0)} = \{(0.5)(0.5)\}^{1/2} = 0.5$ and $\rho_{01} = \text{sgn}(\lambda_{Y0} \cdot \lambda_{M1}) R_{Y(0)} \cdot R_{M(1)} = \{(0.5)(0.6)\}^{1/2} \approx 0.55$. The plausible ranges for the two sensitivity parameters are, for $\rho_0$, [0.1, 0.7], and for $\rho$, [0.1, 0.75]. We see that our elicitation provides a larger value for the correlation of $Y(0,m)$ with $M(1)$ than with $M(0)$, even though the former pair involves different exposure statuses. Our explication of the correlations above show how this occurs. The key in this case is that it was felt that the confounder (financial means) may have a stronger relationship with the oral hygiene index (OHI)

if MomEd is low ($X$=1) than with OHI if MomEd is high ($X$=0), as high mother education may dampen the impact of situational factors, such as financial means (as well as the other potential confounders) on OHI.

Finally, we note a relationship between sensitivity parameters from the copula and hybrid models. Supposing a linear model for each approach, we equate the 'selection bias' terms (of $D$(1), say) obtained for the two approaches in the linear case:

$$-\rho_0 \left( \frac{\sigma_{Y_0}}{\sigma_{M(0)}} \right) \gamma_1 = \beta_1 \phi \,. \qquad \text{(E2)}$$

Note that the left hand side (equal to $-\eta_0 \cdot \gamma$) is the 'bias' from the copula model (as easily obtained from the expression for $D$(1) given in Section 3.1, also (B7) above). From (E2), and using $\sigma^2_{Y_0} = \sigma^2_{Y_0|M(0)} / (1-\rho_0{}^2)$ as suggested in Appendix B, we can solve for $\rho_0$ to obtain,

$$\rho_0 = \text{sgn}\left\{ -\phi \left( \frac{\beta_1}{\gamma_1} \right) \right\} \left( \frac{A_0{}^2}{1-A_0{}^2} \right)^{1/2} \qquad \text{(E3)}$$

where $$A_0 = -\phi \left( \frac{\beta_1}{\gamma_1} \right) \left( \frac{\sigma_{M(0)}}{\sigma_{Y_0|M(0)}} \right).$$

We note that under previous assumptions (including randomization of $X$), the parameters on the right hand side of (E3), apart from $\phi$, are estimable. Thus a value for $\rho_0$ may be obtained, once estimates are substituted for the other parameters, from an elicited value for $\phi$. A similar derivation can be used to provide a possibly different value for $\rho_1 \equiv \rho_{11} \equiv \text{corr}\{Y^*(1,m), M^*(1)\}$ yielding,

$$\rho_1 = \text{sgn}\left\{ -\phi \left( \frac{\beta_1}{\gamma_1} \right) \right\} \left( \frac{A_1{}^2}{1-A_1{}^2} \right)^{1/2} \qquad \text{(E4)}$$

where $A_1 = -\phi\left(\frac{\beta_1}{\gamma_1}\right)\left(\frac{\sigma_{M(1)}}{\sigma_{Y_1|M(1)}}\right)$

Note that similar expressions are not available for $\rho_{01}$ and $\rho_{10}$ as these parameters are not involved in the linear case.

The above expressions could be used to elicit values for $\rho_0$ and $\rho_1$ from $\phi$ (or using the inverse relationship, $\phi$ from $\rho_0$ or $\rho_1$). However, a more useful purpose may be to allow a calibration between the two methods. For example, if an elicited value of $\phi$ corresponds (using (E3) and (E4), respectively) to a larger (or smaller) value for $\rho_0$ and/or $\rho_1$ than that considered as reasonably 'plausible', then one may wish to reconsider the specified confounders and/or confounders effects leading to the elicitation of $\phi$. Similarly, we may wish to inspect the value for $\phi$ calculated as a function (using (E3) or (E4)) of elicited values for $\rho_0$ and/or $\rho_1$.

We illustrate this approach to 'calibration' using the dental data. As in the discussion for the hybrid model, we use linear models for DMFTD and OHI to get estimates for the parameters involved in (E3) and (E4). In the present data, it seems reasonable to assume homogeneity of variances for the two exposure groups so that $\rho_0 = \rho_1$. (Of course, we may examine this assumption itself by comparing the present specifications with those using the above approach for the copula model.) Then, using the plausible ('pessimistic') value from the hybrid model of $\phi$ = -0.4, and plugging in estimates ( $\hat{\beta}_1 = 0.12$ , $\hat{\gamma}_1 = 0.12$ , $\hat{\sigma}_{Y(0)|M(0)} = 0.48$ , $\hat{\sigma}_{M(0)} = 0.78$ ) gives $\rho_0 = \rho_1 \approx 0.28$. This correlation is considerably lower than those (0.5 and 0.55, respectively) obtained from the copula model ('pessimistic') specifications, suggesting that the latter are more pessimistic than that specified from the corresponding hybrid model.

Appendix F. Simulation Study Results on Linear Approximation in Elicitation

In this appendix we present the results of a simulation study intended to give an idea of the accuracy of the linear approximation used in the hybrid model sensitivity parameter elicitation. In this small (single replicate) study, we generated simulated data under the hybrid model for a dichotomous outcome, $Y$, using a logit link and parameter values mimicking our dental data; in addition, a linear (identity-link) regression model was assumed for mediator, $M$. Specifically, for each scenario considered, we generated a data set consisting of $X$, $M$ , and two versions of $Y$: 1) a potential outcome ($Y_C \equiv Y(x,m)$, choosing $x$=0 and $m$ =1 for simplicity and relevance to the mediation estimator) from the hybrid model, to allow estimation of the causal (unbiased) direct effect of $X$; and 2) a variable $Y$ generated from a logistic regression association model (corresponding to the hybrid model with $x = x'$), as would be observed in practice.

The parameters values used, mimicking the dental data, were: $\beta_0 = -1$, $\beta_1 = 0.5$, $\beta_2 = 0.7$, $\gamma_0 = 1$, $\gamma_1 = 0.3$, $\sigma_M = 0.7$ (using the notation of Model (4.1)). The exposure, $X$ , was generated as Bernoulli with $P(X=1) = 0.5$. Different scenarios were obtained using varying sample sizes (200, 2000, and 20,000) and varying values for $\phi$ (namely, -1, -0.4, and 0.4).

Estimation based on the simulated data proceeded by fitting, in turn, the (assumed correctly specified logistic regression) association model and the hybrid model (fixing $m$=1) to the data with the observable outcome, $Y$, and the potential outcome, $Y_C$, respectively, as the final outcome. The estimate of $\phi$ was then obtained as the ratio of the estimated coefficients of $X$ from the hybrid and association models. To study the approximation in question, we fit the same data to linear (identity link) versions of the above association and hybrid models, and the estimate of $\phi$ was obtained as the same ratio as indicated above. We note that the above procedure was chosen in lieu of an approach involving an assumed unobserved confounder ($U$ in our notation in

Appendix E) as it is not straightforward to construct a model consistent with the nonlinear hybrid model that directly involves an unobserved confounder.

Table F1. Simulation estimates of $\phi$ under logistic and linear hybrid models for varying true $\phi$ and sample size ($n$)

| $n$ | 200 | | 2,000 | | 20,000 | |
|---|---|---|---|---|---|---|
| true $\phi$ | logit | lin | logit | lin | logit | lin |
| -1.0 | -0.75 | -0.69 | -0.88 | -0.85 | -1.17 | -1.15 |
| -0.4 | -0.93 | -0.90 | -0.40 | -0.41 | -0.41 | -0.42 |
| 0.4 | 0.16 | 0.16 | 0.61 | 0.63 | 0.41 | 0.43 |

Table F1 provides the estimates of $\phi$ for both the (correct) logistic ("logit") and linear ("lin") versions of the hybrid model. The results show that the estimates of $\phi$ based on the linear approximation are quite close to those from the logistic regression model in each scenario. Though the present study was not intended to evaluate properties of estimators of $\phi$, it should be noted that the results also show that the estimates, even based on the correct logistic regression model, appear to be rather biased in some cases, particularly, for the relatively small sample size of $n$=200. This suggests that one should be cautious in using estimates from small sample studies for the needed parameters in the above elicitation. As also seen in Table F1, the estimators tend to improve as the sample size increases, indicating the consistency of the logistic regression model estimates.

In summary, our simulation study results suggest that estimates of $\phi$ from an assumed linear version of the hybrid model may provide good approximations to those from the logistic regression hybrid model. The results of this study thus supports the use of linear model approximations (as outlined in Appendix E) in the elicitation of values of $\phi$. The reasonableness

of the linear approximation is further suggested by the fact that the resulting errors are likely to be small relative to the uncertainty involved in the specification of unobserved confounder effects. Of course, the approximations will depend on the true parameter values (e.g., coefficients of *X* and *M*) and may be worse or better in other situations relative to the present example; however, we expect the previous comments to pertain fairly generally.

We expect that the linear approximation may be reasonable for other hybrid models as well. For example, we would expect linear approximations of $\phi$ in a hybrid model involving a log link function for a Poisson-distributed count outcome to be at least as good as those shown above for the logistic regression model, as the linear approximation is likely to be better for a count than a dichotomous outcome.

## Appendix G. Additional Data Examples

Here we provide additional examples of the application of the proposed sensitivity analysis methods to the dental data. In these examples, we use the original DMFT (decay, missing, and filled teeth) count rather than dichotomized version. As a first example, we assume a linear (identity link) model, and normal distributions, for both DMFT and OHI. In the second example, we assume DMFT to be distributed as negative binomial and to follow a log-linear model. For both models, we control for the same set of covariates as before. Thus the mediator (*M* = OHI) model is the same as before (namely, (4.1b)), while the alternative (linear and log-linear) *Y* models may be written as,

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 L + \varepsilon_Y \,, \qquad \text{(G1)}$$

$$\log\{E(Y \mid X, L, M)\} = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 L \,, \qquad \text{(G2)}$$

respectively, where $\varepsilon_Y \mid X, M, L \sim N(0,\sigma_Y^2)$, $\sigma_Y^2$ is an unknown variance parameter, and the $\beta$s are unknown parameters (and unrelated for different models, though the same symbols are used for convenience).

The plots for the linear $Y$ model are given in Figure G1 (for the copula and hybrid model approaches) and Figure G2 for the Imai et al. (2010b), approach. The overall ranges for the estimates of $D(1)$ and $I(0)$ across sensitivity parameter values are similar between the Imai et al. and copula approaches. At this point, we comment on theoretical differences between the two approaches. The Imai et al. approach provides estimated natural indirect effects (assuming $I(0) = I(1)$) as a function of a sensitivity parameter, which we denote as $\rho^*$, as follows (in the present notation and after some manipulation):

$$I(0) = \gamma_1 \left( \beta_2 - \frac{\sigma_Y}{\sigma_M} \rho^* \left( \frac{1-\rho_{YM}{}^2}{1-\rho^{*2}} \right)^{\frac{1}{2}} \right)$$

where $\rho_{YM}$ is the estimable correlation between the error for the $M$ model and the error for a (linear) $Y$ model with $\underline{M}$ removed as a covariate. This expression for $I(0)$ has the same form as the formula (B8) from the copula model after the substitution,

$$\rho_0 = \rho^* \left( \frac{1-\rho_{YM}{}^2}{1-\rho^{*2}} \right)^{\frac{1}{2}} .$$

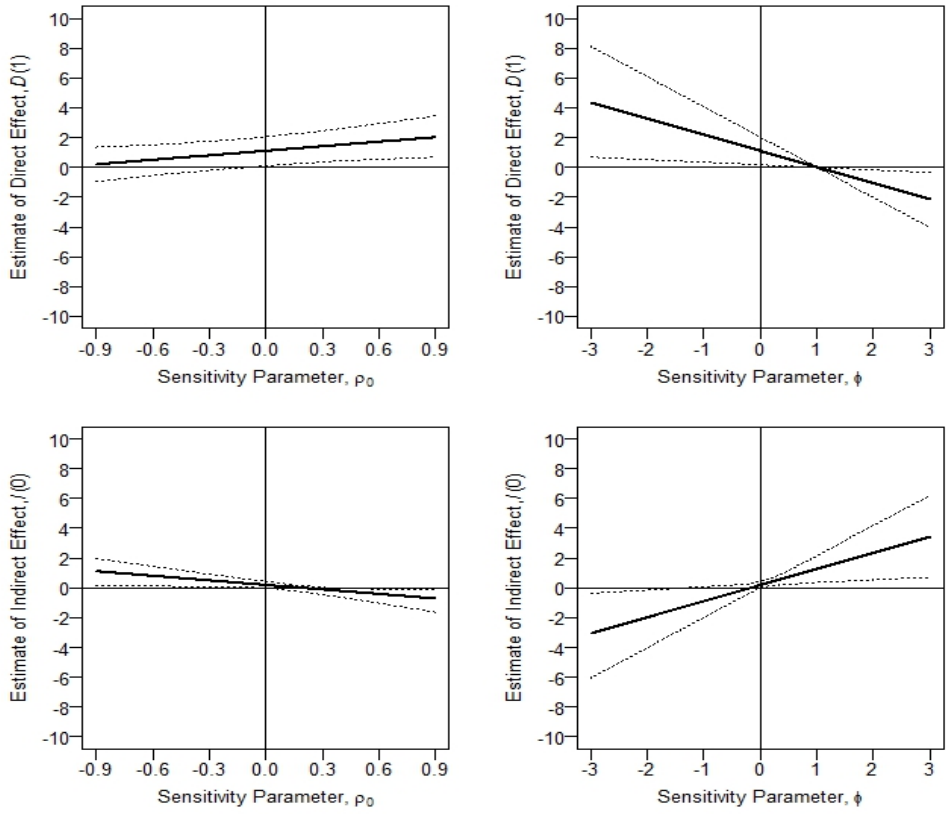


Figure G1. Sensitivity analysis for dental data using a linear *Y* model. Maximum likelihood estimates of direct (top) and indirect (bottom) effects versus sensitivity parameters from copula model (left) and hybrid model (right). Solid line = estimates, Dotted lines = 95% confidence interval bounds.

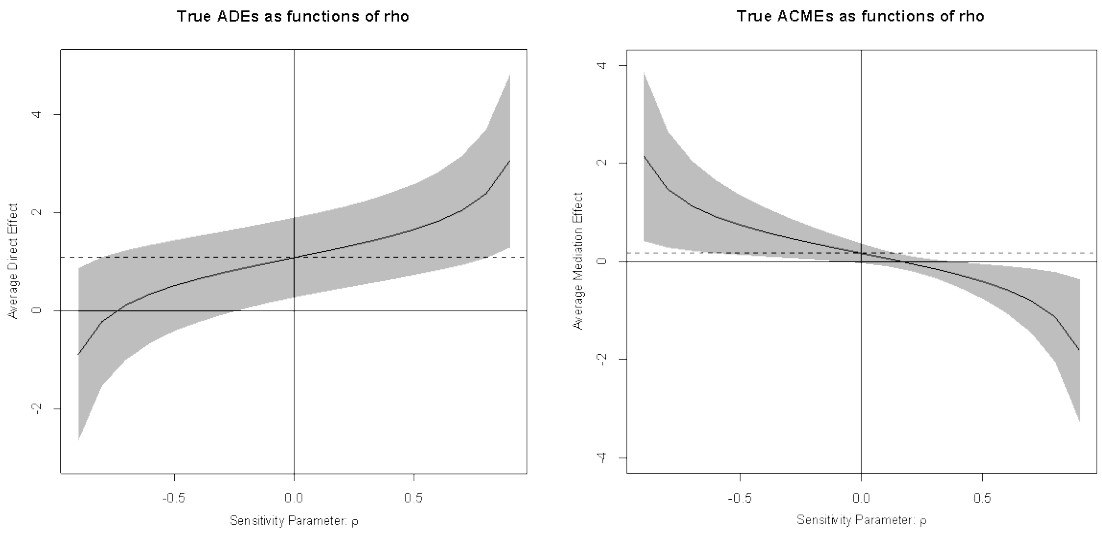


Figure G2. Sensitivity analysis for dental data using the Imai et al. (2010b) approach (linear *Y* and *M* models). The plot on the left (right) is for the natural direct (and indirect) effect, respectively. The solid line corresponds to the estimates, while the shaded area provides 95% confidence interval bounds. (Plots are obtained from the Mediation R package, Tingley et al.)

From the above results, we note the following differences between the Imai et al. and the copula method approach:

a) The two approaches have difference sensitivity parameters ($\rho$ versus $\rho^*$), resulting from different models. (Note that the Imai et al. approach involves a model for *Y* leaving *M* out as a covariate.)

b) In general, a range of [-1,1] for one of these parameters implies a different range (possibly including values outside of [-1,1]) for the other. Thus, the two implied models are not generally compatible.

c) The sensitivity parameters have different interpretations as indicated by their definitions. Imai et al.: $\rho^* = \text{corr}\{M(X_i), \varepsilon_Y(X_i,M(X_i))\}$ where $\varepsilon_Y(X_i,M(X_i)) = Y(X_i,M(X_i)) - \beta_0 - \beta_1 X_i - \beta_2 M(X_i) - \beta_3 L_i$; copula model: $\rho_0 = \text{corr}\{Y(0,m),M(0)\}$. Thus, $\rho^*$ is the correlation between, for example, $M(0)$ and the potential outcome of the residual of $Y$ given $X$=0 and the observed value $M(0)$, while $\rho_0$ represents the correlation between the potential mediator ($M(0)$) and the potential outcome (equivalently, the residual) of $Y$ at $X$=0 and a fixed value ($m$) for $M$.

Next we illustrate the two new sensitivity analysis approaches (again using the dental data with the DMFT count as the final outcome, $Y$) assuming $Y$ is distributed as negative binomial and following a log-linear model. For the copula model, we chose to reduce the parameterization by specifying $\rho$ from $\rho_0$ using the relationship, $\log\{\rho/(1-\rho)\} = \log\{\rho_0/(1-\rho_0)\} + \log c$, equivalently, $\rho = c\,\rho_0 / \{1 + (c\text{-}1)\,\rho_0\}$. We used $c$ = 1.2 which yields similar values as those elicited for the copula model in Appendix E. The plots of the estimates (for $D(1)$ and $I(0)$) versus the respective sensitivity parameters for the copula and hybrid model approaches are shown in Figure G3. While there are differences, particularly in the confidence bands, the overall patterns for the mediation effect estimates are similar for the different models for DMFT.

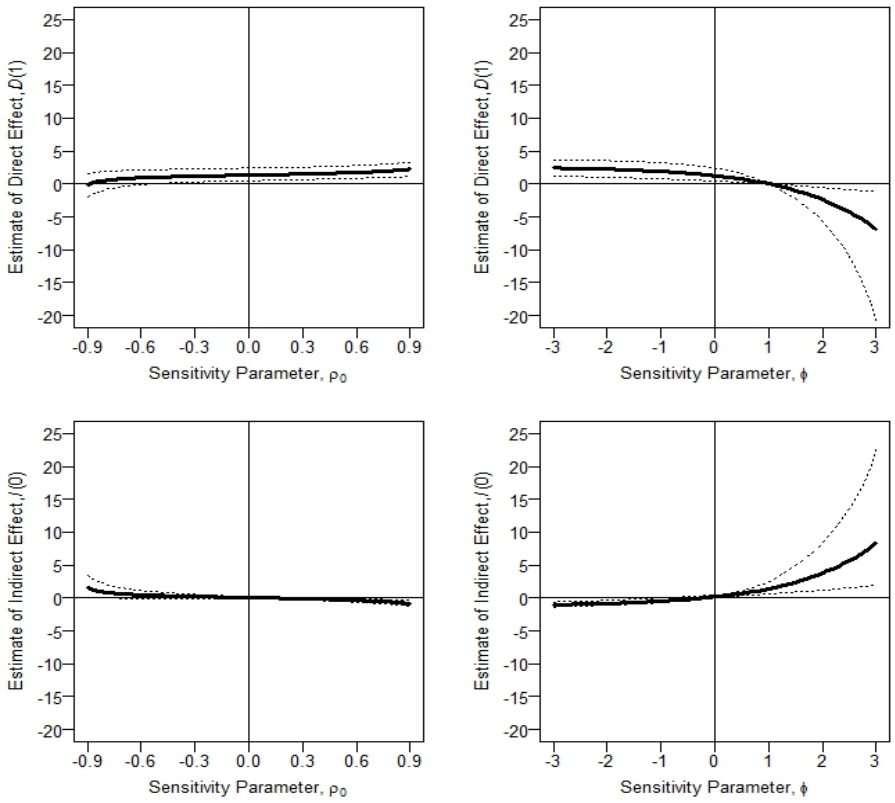


Figure G3. Sensitivity analysis for dental data using a loglinear *Y* model. Maximum likelihood estimates of direct (top) and indirect (bottom) effects versus sensitivity parameters from copula model (left) and hybrid model (right). Solid line = estimates, Dotted lines = 95% confidence interval bounds.

Additional reference for Supplementary Material

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, Forthcoming.