Accepted Manuscript. Final article published in *Structural Equation Modeling: A Multidisciplinary Journal*. Free copy here while lasts. doi: 10.1080/10705511.2015.1062730

Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention

Trang Quynh Nguyen,[a] Yenny Webb-Vargas,[b] Ina M. Koning[c] and Elizabeth A. Stuart[a,b,d]
[a]Department of Mental Health, [b]Department of Biostatistics, [d]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health;
[c]Department of Interdisciplinary Social Science, University of Utrecht

We investigate a method to estimate the combined effect of multiple continuous/ordinal mediators on a binary outcome: 1) fit a structural equation model with probit link for the outcome and identity/probit link for continuous/ordinal mediators, 2) predict potential outcome probabilities, and 3) compute natural direct and indirect effects. Step 2 involves rescaling the latent continuous variable underlying the outcome to address residual mediator variance/covariance. We evaluate the estimation of risk-difference- and risk-ratio-based effects (RDs, RRs) using the ML, WLSMV and Bayes estimators in Mplus. Across most variations in path-coefficient and mediator-residual-correlation signs and strengths, and confounding situations investigated, the method performs well with all estimators, but favors ML/WLSMV for RDs with continuous mediators, and Bayes for RRs with ordinal mediators. Bayes outperforms WLSMV/ML regardless of mediator type when estimating RRs with small potential outcome probabilities and in two other special cases. An adolescent alcohol prevention study is used for illustration.

Keywords: causal mediation analysis, binary outcome, multiple mediators, ordinal mediators, continuous mediators, structural equation modeling

The causal inference approach has added tremendously to the mediation analysis literature, through defining direct and indirect effects with causal interpretation, clarifying assumptions that allow their identification, and advancing procedures to estimate them in different settings (e.g., Albert, 2012; Coffman & Zhong, 2012; Huang, Sivaganesan, Succop, & Goodman, 2004; Imai, Keele, & Yamamoto, 2010; Pearl, 2001, 2009; Petersen, Sinisi, & van der Laan, 2006; Robins & Greenland, 1992; Ten Have & Joffe, 2012; VanderWeele & Vansteelandt, 2009, 2010). The last few years have seen many applications of causal mediation analysis, especially in epidemiology (e.g., Ananth & VanderWeele, 2011; Bennett, Rankin, & Rosenberg, 2012; Nandi, Glymour, Kawachi, & VanderWeele, 2012; Smith, Smith, Mustard, Lu, & Glazier, 2013; Subbaraman, Lendle, van der Laan, Kaskutas, & Ahern, 2013). Recently, methods using structural equation modeling (SEM) to estimate causally defined effects were introduced (B. O. Muthén, 2011; B. O. Muthén & Asparouhov, 2015), making these effect definitions more accessible to social scientists familiar with SEM.

Many social and behavioral studies examine multiple mediating pathways. For example, studying the relationship between childhood abuse/neglect and alcohol abuse, Schuck and Spatz (2001) consider a range of psychological constructs as mediators, including depression, isolation/loneliness, worthlessness, low self-esteem and endorsement of alcohol/drug use as a coping strategy. Kelly, Hoeppner, Stout, and Pagano (2012) assess both individual (self-efficacy, depression and spiritual/religious practice) and social (pro-abstinence and pro-drinking social networks) factors as mediators of the effect of Alcoholics Anonymous attendance on alcohol

Correspondence should be addressed to Trang Quynh Nguyen, 624 N. Broadway, Rm. 896, Baltimore, MD 21205; email: nqtrang.hanoi@gmail.com.

TQN was supported by NIDA grant T-32DA007292 (PI: C.D.M. Furr-Holden). EAS was supported by NIMH grant R01MH099010. YWV was supported by NIBIB grant 5R01EB016061-02 (PI: M. Lindquist). The alcohol prevention study was funded by the Dutch Health Care Research Organization, grant 6220.0021. The authors thank the anonymous reviewers for their helpful critique.

This paper's Web appendices can be found at http://trang-q-nguyen.weebly.com/methods-papers.html.

consumption and abstinence. Koning, van den Eijnden, Engels, Verdurmen, and Vollebergh (2010) evaluate a school-based intervention that targeted adolescent self-control and attitudes about alcohol, as well as parents' rules and attitudes concerning adolescent alcohol use, in order to reduce adolescent drinking.

These multiple-mediator studies have generally used the traditional association approach, which was started by Baron and Kenny (1986) for a single mediator, with methods later developed for multiple-mediator situations (see MacKinnon, 2008, Ch. 5-6). To date, the causal mediation methodological literature has largely focused on the single mediator case. Recent work starts to extend causal inference methods to multiple-mediator situations, with a key theme being the mediators' dependence structure. It has been noted that the (often implicit) assumption commonly made in applied multiple-mediator analysis, that the mediators are independent conditional on the exposure and baseline covariates, is unrealistic and when violated leads to biased estimates of causal effects (Imai & Yamamoto, 2013; VanderWeele & Vansteelandt, 2013). Work by Imai and Yamamoto (2013) tackles the problem of identifying the effect through a mediator of interest that is causally affected by other mediators. Also with causally related mediators, Daniel, De Stavola, Cousens, and Vansteelandt (2015) show the complexity of total effect decomposition, with alternatives using different sets of effects through different path combinations.

Estimation of the combined effect of multiple mediators is a simpler objective that is also useful. It has been tackled by VanderWeele and Vansteelandt (2013), who propose several regression-based methods that handle continuous outcomes (using linear regression), rare binary outcomes (using logistic regression plus an approximation in deriving causal effects), and non-rare binary outcomes (using log-linear regression), as well as a weighting-based method. Our paper considers a different regression-based method to address the combined effect of multiple continuous or ordinal mediators when the outcome is binary – extending an existing method using probit SEM to estimate natural direct and indirect effects for a binary outcome and a single mediator (B. O. Muthén, 2011). As this method does not require a rare outcome like the above logistic regression-based method and is less likely to have convergence problems than log-linear regression, it is a potentially useful addition to the current method choices for dealing with binary outcomes.

Real data example. In The Prevention of Alcohol Use in Students (PAS) trial in the Netherlands, middle schools were randomized to one of four conditions: student intervention (promoting healthy attitudes and strengthening refusal skills), parent intervention (encouraging parental rule setting), student-and-parent combined intervention, and control condition (regular biology curriculum covering effects of alcohol). The combined intervention was effective in reducing drinking onset (Koning, Van Den Eijnden, Verdurmen, Engels, & Vollebergh, 2011; Koning et al., 2009) and drinking frequency (Koning et al., 2009). Examining variables targeted by the intervention, a (non-causal) mediation analysis found that adolescent self-control, adolescent attitudes about alcohol, and adolescent-reported parental rules about alcohol mediated the relationship between the combined intervention and onset of weekly drinking by 22 months follow-up (Koning et al., 2010). In addition to identifying likely mediators, there is generally also an interest in inferring causality and estimating the mediated effect, for example, in terms of reduction in drinking. For these purposes the proposed method is useful. To illustrate, we conduct a similar analysis restricted to the parent-and-student combined intervention versus control, with the outcome being weekly drinking at 22 months. We consider the same hypothesized mediators, and partition the total intervention effect (reduction in drinking prevalence, comparing intervention condition to control condition) into: (i) an effect mediated by the mediators (conceptualized as reduction in drinking prevalence comparing the intervention condition to a hypothetical condition of intervention participation where the intervention's effects on the mediators are blocked) and (ii) an unmediated effect (comparing this hypothetical condition to the control condition). These, in the causal mediation literature, are called *natural indirect* and *natural direct effects* (Pearl, 2001).

In the next sections we present the proposed method, including a formal definition of natural direct and indirect effects based on the potential outcome framework (Rubin, 1974), the model used, identifying assumptions, and estimation procedures. We report results from simulation studies before applying the method to the above example. We conclude with recommendations for application and discussion of future research. Mplus inputs and R code for method implementation are included in the Web appendices.

Of the three common measures of effect on a binary outcome, the risk difference (RD), risk ratio (RR) and odds ratio (OR), also called absolute risk, relative risk and relative odds (Rothman, Greenland, & Lash, 2008), in this study we consider the RD and RR. There are reasons to suspect the proposed method works less well with the OR, but it may be appropriate in certain cases, which we mention in the Discussion section.

The proposed method

Definition of causal mediation effects: a review

Consider a binary exposure X, a binary outcome Y, and k mediators of the $X \rightarrow Y$ relationship, $M^{[1]}, M^{[2]}, ..., M^{[k]}$ contained in vector $\boldsymbol{M}$. For person i, the potential values of the mediators if the exposure were to take the value x is denoted by $\boldsymbol{M}_i(x)$. With the binary exposure, there are two potentials sets of mediator values, $\boldsymbol{M}_i(0)$ and $\boldsymbol{M}_i(1)$, of which only one happens; the other is contrary to fact. Also for person i, the potential outcome if the exposure were to take the value x is denoted by $Y_i(x)$; there are two such potential outcomes $Y_i(0)$ and $Y_i(1)$, one of which is contrary to fact. The potential outcome if the exposure were to take the value x and if the mediators were to take the values $\boldsymbol{m}$ is denoted by $Y_i(x, \boldsymbol{m})$. With continuous mediators, there are an infinite number of $(x, \boldsymbol{m})$ combinations, and for each person, all but one are contrary to fact.

Natural direct and *indirect effects* are defined based on a special set of potential outcomes, denoted by $Y_i(x, \boldsymbol{M}_i(x'))$: potential outcome if the exposure were to take the value x AND the mediators were to take the values that they would take if the exposure were to take the value x'. x and x' could be either 0 or 1 and may or may not be the same. Each person has four such special potential outcomes: $Y_i(0, \boldsymbol{M}_i(0))$, $Y_i(1, \boldsymbol{M}_i(1))$, $Y_i(1, \boldsymbol{M}_i(0))$, and $Y_i(0, \boldsymbol{M}_i(1))$. Of these, the first two are partially observed – $Y_i(0, \boldsymbol{M}_i(0))$ observed for those with $X_i = 0$, and $Y_i(1, \boldsymbol{M}_i(1))$ observed for those with $X_i = 1$. The latter two are useful for defining mediation effects, but are completely hypothetical (or 'truly counterfactual'), because $X_i = 1$ and $\boldsymbol{M}_i = \boldsymbol{M}_i(0)$ do not co-occur, and neither do $X_i = 0$ and $\boldsymbol{M}_i = \boldsymbol{M}_i(1)$. (There are criticisms of the use of these truly counterfactual quantities, which are outside the scope of this paper – see Rubin, 2004, for example.)

The expected values of the potential outcomes $Y_i(x, \boldsymbol{M}_i(x'))$ over the population are potential outcome probabilities, $\mathrm{P}[Y(x, \boldsymbol{M}(x')) = 1]$, which serve as the basis for defining causal effects. To simplify notation, we use $p_{xx'}$ as an abbreviation for $\mathrm{P}[Y(x, \boldsymbol{M}(x')) = 1]$. In $p_{xx'}$, the p component denotes the probability that the potential outcome is 1, the x index refers to a (possibly contrary to fact) exposure condition, and the x' index refers to a set of (possibly contrary to fact) mediator values that correspond to exposure condition x'.

On the risk difference scale, the total effect (TE) is:

$$TE_{RD} = p_{11} - p_{00}.$$

A direct effect is an effect on the outcome of changing the exposure, say from 0 to 1, but blocking any effect this might have on the mediators. There are two such effects setting the mediators at the levels they would be for exposure condition 0 and for exposure condition 1, both called *natural direct effects* (NDE), which we denote $NDE(\cdot 0)$ and $NDE(\cdot 1)$,

$$NDE(\cdot 0)_{RD} = p_{10} - p_{00};$$
$$NDE(\cdot 1)_{RD} = p_{11} - p_{01}.$$

An indirect effect is an effect on the outcome of changing the mediators as if by changing the exposure from 0 to 1, but at the same time fixing the exposure at one value. There are two such effects setting the exposure at 0 and setting the exposure at 1, both called *natural indirect effects* (NIE), which we denote $NIE(0\cdot)$ and $NIE(1\cdot)$,

$$NIE(0\cdot)_{RD} = p_{01} - p_{00};$$
$$NIE(1\cdot)_{RD} = p_{11} - p_{10}.$$

Note that TE can be decomposed into NDE and NIE in two different ways:

$$TE_{RD} = NDE(\cdot 0)_{RD} + NIE(1\cdot)_{RD};$$
$$TE_{RD} = NIE(0\cdot)_{RD} + NDE(\cdot 1)_{RD},$$

Similarly, on a risk ratio scale, TE, NDE and NIE are defined as:

$$TE_{RR} = \frac{p_{11}}{p_{00}};$$

$$NDE(\cdot 0)_{RR} = \frac{p_{10}}{p_{00}};\ NDE(\cdot 1)_{RR} = \frac{p_{11}}{p_{01}};$$

$$NIE(0\cdot)_{RR} = \frac{p_{01}}{p_{00}};\ NIE(1\cdot)_{RR} = \frac{p_{11}}{p_{10}};$$

and the TE decompositions are:

$$TE_{RR} = NDE(\cdot 0)_{RR} \times NIE(1\cdot)_{RR};$$
$$TE_{RR} = NIE(0\cdot)_{RR} \times NDE(\cdot 1)_{RR}.$$

Identification of effects

If we observed all four $Y_i(x, \boldsymbol{M}_i(x'))$ for everyone in the sample, we could compute TE, NDE and NIE by averaging these quantities over the sample to get potential outcome probabilities and taking the differences or ratios of relevant pairs of probabilities. However, only one potential outcome is observed for each individual in the sample, and two cannot be observed at all. In order for the potential outcome probabilities to be identified based on the data, we would have to be willing to make some assumptions.

The first assumption for this method is a causal model for the potential mediator values and potential outcomes (assumption 0). We assume linear models for continuous, and probit models for ordinal, mediators; and a probit model for the outcome. With an ordinal mediator, we assume that the actual mediator is an unobserved continuous variable underlying the ordinal variable (this

Table 1
Observed and unobserved potential mediator values and potential outcomes

Exposure	$\boldsymbol{M}_i(0)$	$\boldsymbol{M}_i(1)$	$Y_i(0, \boldsymbol{M}_i(0))$	$Y_i(1, \boldsymbol{M}_i(1))$	$Y_i(1, \boldsymbol{M}_i(0))$	$Y_i(0, \boldsymbol{M}_i(1))$
$X_i = 1$	✗	✓	✗	✓	✗✗	✗✗
$X_i = 0$	✓	✗	✓	✗	✗✗	✗✗

✓: observed; ✗: not observed; ✗✗: not observed and truly counterfactual

method thus does not apply to situations where this is conceptually unlikely or implausible). The parameters of this causal model, if known, would allow us to compute potential outcome probabilities (see details shortly).

To use observed data to infer the parameters of this causal model, however, we need the following five identifying assumptions. Four of these (assumptions 1–4) have been thoroughly discussed in the causal mediation literature for the case of a single mediator (see VanderWeele & Vansteelandt, 2009, for example). Assumption 5 is needed for this particular method of dealing with multiple mediators.

1. no unmeasured exposure-mediator confounding
2. no unmeasured exposure-outcome confounding
3. no unmeasured mediator-outcome confounding
4. no mediator-outcome confounder that is influenced by the exposure
5. no mediator-mediator interaction in influencing the outcome

With these assumptions, our assumed model can be represented as follows. Here we add the notation of $\boldsymbol{Z}$, a vector of all confounders of any of the $X \to \boldsymbol{M}, X \to Y$, $\boldsymbol{M} \to Y$ relationships (see Figure 1), all elements of $\boldsymbol{Z}$ are observed (assumptions 1–3), and none is causally influenced by X (assumption 4). For a person i, the potential values of the mediators if the exposure were to take the value x are determined by

$$\boldsymbol{M}_i(x) = \boldsymbol{\mu_M} + \boldsymbol{\alpha} x + \boldsymbol{\Lambda} \boldsymbol{z}_i + \boldsymbol{\epsilon}_{\boldsymbol{M}i},$$

where vector $\boldsymbol{\alpha}$ represents the effects of the exposure, and matrix $\boldsymbol{\Lambda}$ represents the effects of the confounders on the mediators; vector $\boldsymbol{z}_i$ contains the person's own confounders; and $\boldsymbol{\mu_M}$ is a vector of intercepts. The error vector $\boldsymbol{\epsilon}_{\boldsymbol{M}i}$ reflects influences on the mediators for this person that are independent of confounders and exposure condition. Note that a person's potential mediator values are partly determined by his/her confounders. To simplify notation, we can drop the subscript i and rewrite this as a population model:

$$\boldsymbol{M}(x)|\boldsymbol{z} = \boldsymbol{\mu_M} + \boldsymbol{\alpha} x + \boldsymbol{\Lambda} \boldsymbol{z} + \boldsymbol{\epsilon_M}, \tag{1}$$

where $\boldsymbol{z}$ is any pattern of the confounders that exists in the population. The error terms $\boldsymbol{\epsilon_M}$ are distributed multivariate normal with mean $\boldsymbol{0}$ and covariance $\boldsymbol{\Sigma}$. The off-diagonal elements of $\boldsymbol{\Sigma}$ may be non-zero, meaning that the mediators may be correlated due to reasons other than the influence of the exposure and confounders.

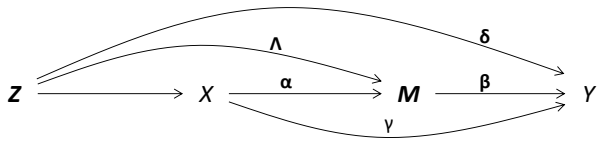

Figure 1. Diagram representing the causal model

With ordinal mediator variables, we use probit models, assuming that the ordinal variables $\boldsymbol{M}$ represent underlying latent continuous variables $\boldsymbol{M}^*$ that relate to X and Z via the linear model defined by equation 1, and the manifest $\boldsymbol{M}$ relates to the latent $\boldsymbol{M}^*$ via sets of thresholds $\boldsymbol{\tau_M}$ that split each continuous latent mediator into ordinal categories. In this case, the intercepts $\boldsymbol{\mu_M}$ are set to $\boldsymbol{0}$, the variance of each error term in $\boldsymbol{\epsilon_M}$ is fixed at 1, and $\boldsymbol{\Sigma}$ is a correlation matrix.

For the outcome Y, we also use a probit model, which is equivalent to a normal linear model for a latent continuous variable Y^* underlying the binary Y, with a threshold τ_Y, i.e., $Y = 1$ if $Y^* > \tau_Y$ and 0 otherwise. The potential outcomes if the exposure were to take the value x and the (continuous) mediators were to take the values $\boldsymbol{m}$ are determined by

$$Y^*(x, \boldsymbol{m})|\boldsymbol{z} = \gamma x + \boldsymbol{\beta}^T \boldsymbol{m} + \boldsymbol{\delta}^T \boldsymbol{z} + \epsilon_Y, \tag{2}$$

where γ, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ denote the effects of the exposure, mediators and confounders on the outcome. The error term ϵ_Y reflects influences on the outcome that are independent of the exposure, confounders and mediators. ϵ_Y is distributed normal mean 0 and variance 1, and is independent of $\boldsymbol{\epsilon_M}$. With ordinal mediators, $\boldsymbol{m}$ in equation 2 represents values of the latent continuous $\boldsymbol{M}^*$ underlying $\boldsymbol{M}$.

Under assumptions 1–4, the model parameters $\boldsymbol{\mu_M}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \gamma, \boldsymbol{\beta}, \delta, \boldsymbol{\Sigma}, \tau_Y$ can be identified via regression analysis with observed data, using a model that replaces the potential mediators and potential outcomes with their observed counterparts.

$$\boldsymbol{M}|x, \boldsymbol{z} = \boldsymbol{\mu_M} + \boldsymbol{\alpha} x + \boldsymbol{\Lambda} \boldsymbol{z} + \boldsymbol{\epsilon_M},$$
$$Y^*|x, \boldsymbol{m}, \boldsymbol{z} = \gamma x + \boldsymbol{\beta}^T \boldsymbol{m} + \boldsymbol{\delta}^T \boldsymbol{z} + \epsilon_Y.$$

With $\boldsymbol{\mu_M}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \gamma, \boldsymbol{\beta}, \delta, \boldsymbol{\Sigma}, \tau_Y$ identified, TE, NDE and NIE are also identified, because the potential outcome probabilities $p_{xx'}$ are functions of these parameters. Replacing $\boldsymbol{m}$ in equation 2 with the quantity for $\boldsymbol{M}(x')|\boldsymbol{z}$ based on equation 1 gives,

$$\begin{aligned} Y^*(x, \boldsymbol{M}(x'))|\boldsymbol{z} &= \gamma x + \boldsymbol{\beta}^T(\boldsymbol{M}(x')|\boldsymbol{z}) + \boldsymbol{\delta}^T\boldsymbol{z} + \epsilon_Y \\ &= \gamma x + \boldsymbol{\beta}^T(\boldsymbol{\mu_M} + \boldsymbol{\alpha}x' + \boldsymbol{\Lambda}\boldsymbol{z} + \boldsymbol{\epsilon_M}) + \boldsymbol{\delta}^T\boldsymbol{z} + \epsilon_Y \\ &= \boldsymbol{\beta}^T\boldsymbol{\mu_M} + (\boldsymbol{\beta}^T\boldsymbol{\Lambda} + \boldsymbol{\delta}^T)\boldsymbol{z} + \gamma x + \boldsymbol{\beta}^T\boldsymbol{\alpha}x' + \\ &\quad + (\boldsymbol{\beta}^T\boldsymbol{\epsilon_M} + \epsilon_Y), \end{aligned} \tag{3}$$

In this equation, the error combination $(\boldsymbol{\beta}^T\boldsymbol{\epsilon_M} + \epsilon_Y)$ is distributed normal, but its variance, $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + 1$, is greater than 1. Conversion from potential values of Y^* to potential Y probabilities thus involves rescaling the continuous metric so that error variance is equal to 1 before taking the inverse-probit,

$$\mathrm{P}[(Y(x, \boldsymbol{M}(x'))|\boldsymbol{z}) = 1] = \Phi\left[\frac{(-\tau_Y + \boldsymbol{\beta}^T\boldsymbol{\mu_M}) + (\boldsymbol{\beta}^T\boldsymbol{\Lambda} + \boldsymbol{\delta}^T)\boldsymbol{z} + \gamma x + \boldsymbol{\beta}^T\boldsymbol{\alpha}x'}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + 1}}\right]. \tag{4}$$

(Φ is the inverse-probit, or standard normal cumulative distribution, function.)

This method of combining normal error terms from probit/linear models and rescaling the latent variable underlying the binary outcome is an extension of B. O. Muthén (2011), which uses a similar rescaling strategy for a binary outcome and a single mediator; the same rescaling also appears in Imai, Keele, and Tingley (2010). In the single mediator case the rescaling incorporates the mediator's residual variance (after accounting for the influence of the exposure and confounders). With multiple mediators, this is replaced by $\boldsymbol{\Sigma}$, the residual covariance matrix of multiple continuous mediators, or the residual correlation matrix of multiple ordinal mediators. This matrix includes the mediators' residual variances (diagonal elements) and residual covariances/correlations (off-diagonal elements). This feature handles the mediators' residual dependence (dependence not explained by the exposure and confounders) in computing causal mediation effects.

If the effect of a mediator on the outcome may differ by exposure condition (or by confounder level), this can be incorporated by adding interaction terms $x\boldsymbol{m}$ (or $\boldsymbol{z}\boldsymbol{m}$) to equation 2. This would change the terms in equation 3, but the error combinations remain normally distributed which allows using the inverse-probit function to convert to probability. The proposed method requires that there is no mediator-mediator interaction (assumption 5), however, because such an interaction would result in error combinations being non-normal, due to a product of the mediators' error terms.

A simple way to think about equation 4 is to rewrite it as

$$\mathrm{P}[(Y(x, \boldsymbol{M}(x'))|\boldsymbol{z}) = 1] = \Phi\left(\frac{\theta_0 + \boldsymbol{\theta}_z^T\boldsymbol{z} + \theta_x x + \theta_{x'}x'}{\theta_{\text{scale}}}\right). \tag{5}$$

where θ_0 is a constant term, $\boldsymbol{\theta_z}$ is a set of coefficients for the confounders, θ_x is a coefficient for exposure condition x, $\theta_{x'}$ is a coefficient for exposure condition x', and θ_{scale} is a scale parameter; and these "new" parameters are defined as $\theta_0 = -\tau_Y + \boldsymbol{\beta}^T\boldsymbol{\mu_M}$ (or $= -\tau_Y$ with ordinal mediators), $\boldsymbol{\theta}_z^T = \boldsymbol{\beta}^T\boldsymbol{\Lambda} + \boldsymbol{\delta}^T$, $\theta_x = \gamma$, $\theta_{x'} = \boldsymbol{\beta}^T\boldsymbol{\alpha}$, and $\theta_{\text{scale}} = \sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + 1}$. In non-matrix notation, with n confounders $Z_{[1]}, ..., Z_{[n]}$,

$$\mathrm{P}[(Y(x,\boldsymbol{M}(x'))|\boldsymbol{z}) = 1] = \Phi\left(\frac{\theta_0 + \theta_{z_{[1]}}z_{[1]} + ... + \theta_{z_{[n]}}z_{[n]} + \theta_x x + \theta_{x'}x'}{\theta_{\text{scale}}}\right). \tag{6}$$

Note that this equation gives potential outcome probabilities conditional on $\boldsymbol{z}$, not the marginal $p_{xx'} = \mathrm{P}[Y(x, \boldsymbol{M}(x')) = 1]$. In certain cases researchers might choose to use conditional potential outcome probabilities (as implemented in B. O. Muthén, 2011), if the interest is in causal mediation effects for one or more (sets of) values for $\boldsymbol{Z}$, for example the effects among men (or women) with average height. More generally, averaging over the sample distribution of $\boldsymbol{Z}$ provides estimates of the marginal potential outcome probabilities,

$$p_{xx'} = \int \mathrm{P}[(Y(x, \boldsymbol{M}(x'))|\boldsymbol{z}) = 1]f(\boldsymbol{z})\mathrm{d}\boldsymbol{z}. \tag{7}$$

In the special case with no confounding, the potential outcome probabilities are simply,

$$p_{xx'} = \Phi\left(\frac{\theta_0 + \theta_x x + \theta_{x'}x'}{\theta_{\text{scale}}}\right). \tag{8}$$

Estimation procedures

Estimation is based on a three-step procedure: (1) fitting a structural equation model, using probit link for the outcome and identity/probit link for continuous/ordinal mediators, (2) predicting potential outcome probabilities $p_{xx'}$ based on parameter estimates, and (3) computing TE, NDE and NIE by taking differences or ratios of these probabilities.

Model fitting is implemented in Mplus 7.2 (L. K. Muthén & Muthén, 1998-2012). We use three estimators that allow fitting the model in Mplus: ML (with probit link) for continuous mediators; and WLSMV and Bayes for both continuous and ordinal

mediators.[1] For brevity we sometimes refer to them as 'ML', 'WLSMV' and 'Bayes' without saying 'estimator', except where that would be confusing.

Without confounding, $p_{xx'}$, TE, NDE, NIE are functions of model parameters only; they are built into the Mplus input, so the three steps are combined in one. With confounding, these quantities are functions of model parameters and data (confounder values); the model is run in Mplus and $p_{xx'}$, TE, NDE, NIE are computed outside of Mplus.

When using ML and WLSMV, in the no confounding case, Mplus provides Delta method-based confidence intervals for these quantities; in the confounding case, confidence intervals are obtained via bootstrapping from the data (with 500 bootstrap samples). With Bayes estimator, we use non-informative priors and extract median point estimates and quantile credible intervals from the posterior distribution (all Mplus default choices).

Simulation studies

We investigate several situations. Starting with the simplified case of no confounding, we examine method performance across variations in the signs and magnitudes of path coefficients and mediator residual correlations. We then investigate variations in confounding, including confounding of mediator-outcome relationships only (with exposure randomized) and confounding of all paths. We also examine situations with small potential outcome probabilities. For this initial investigation, we consider only cases with no model misspecification, i.e., the data generation model that matches the analysis model.

A three-mediator setup is used throughout, with mediators being either all continuous or all ordinal. With continuous mediators, we set residual variances to 1 and intercepts to 0. This is without loss of generality, because any continuous variable can be converted to this form through a simple rescaling and location shift, both of which do not affect the strength of the relationships among the variables or the causal mediation effects. This choice aligns the (i) the manifest mediators in the continuous mediators setup with (ii) the latent mediators in the ordinal mediators setup (when using the same set of path coefficients and mediator residual correlations), which allows comparing method performance between these two types of mediator variables.

For each scenario investigated, we use 1000 simulations of sample size 500. We track the estimator's bias, standard deviation (SD), root mean squared error (RMSE), estimated standard error (SE), and proportion of the 95% confidence/credible intervals (CIs) that cover the true effect (coverage). For RD-based effects, which share the same raw scale (proportion), we use raw bias, SD, RMSE and SE. For RR-based effects, which are ratios on very different scales, we standardize these quantities by dividing them by the true effects.

Variation in path and correlation signs (with no confounding). We set the absolute values of all path coefficients to 1 and of all mediator residual correlations to 0.4, and vary their signs. Twenty-two scenarios are examined (see Table 2), each combining one of 10 combinations of path signs and one of four mediator residual correlation matrices (one with all three positive correlations, and three each with one negative correlation). These *sign scenarios* represent all permutations of path and correlation signs, in the sense that each permutation that is different from these can be converted to one of them by (i) flipping the sign(s) – or reversing the category order(s) – of certain mediator(s) or of the outcome, and/or (ii) switching mediator locations (see more detail in Web Appendix A).

Variation in path and correlation strengths (with no confounding). We compare uniform path strengths (all with absolute value 1) with two other

[1] Maximum likelihood estimation finds parameter estimates that maximize the likelihood function, which is the joint probability density of model parameters and observed data. In this investigation, we use the simplest maximum likelihood estimator in Mplus, ML, which assumes normality of continuous variables.

Another estimation approach is least squares estimation, which finds parameter estimates that minimize a fit function based on the squared differences between observed and model predicted values. For ordinary least squares estimation, which assumes homoscedastic and uncorrelated data, the fit function is the sum of squares. For weighted least squares estimation, which does not require the same assumptions, the fit function is a weighted sum of squares, using a weight matrix that contains information about variances and covariances of the data. Weighted least squares is commonly used in SEM – see Savalei (2014) for an accessible introduction to this topic. Of the weighted least squares estimators available in Mplus, we use WLSMV, Mplus's default estimator for models with ordinal dependent variables. WLSMV is a three-stage estimator that uses the diagonal weight matrix to obtain estimates and corrects the standard errors and test statistics using the full weight matrix – see B. O. Muthén, du Toit, and Spisic (1997) for technical details.

The Bayes estimator implements Bayesian analysis, a description of which is beyond the scope of this paper. The key concept is that model parameters are assigned prior distributions representing prior beliefs about them; data are used to update these beliefs, which are represented in posterior distributions; and posterior distributions are used to make inference about the parameters. Readers not familiar with Bayesian methods could consult textbooks and courses about this area of statistics. For technical details about the Bayes estimator in Mplus, see Asparouhov and Muthén (2010).

Table 2
Scenarios for variation in path and correlation signs

Scenario	α_1	α_2	α_3	β_1	β_2	β_3	γ	ρ_{12}	ρ_{13}	ρ_{23}
1p	1	1	1	1	1	1	1	0.4	0.4	0.4
1a								−0.4	0.4	0.4
2p								0.4	0.4	0.4
2a	1	1	1	−1	1	1	1	−0.4	0.4	0.4
2c								0.4	0.4	−0.4
3p								0.4	0.4	0.4
3a	1	1	1	−1	−1	1	1	−0.4	0.4	0.4
3b								0.4	−0.4	0.4
4p	1	1	1	−1	−1	−1	1	0.4	0.4	0.4
4a								−0.4	0.4	0.4
5p	−1	1	1	1	1	1	1	0.4	0.4	0.4
5c								0.4	0.4	−0.4
6p	−1	1	1	−1	1	1	1	0.4	0.4	0.4
6c								0.4	0.4	−0.4
7p	−1	1	1	1	−1	1	1	0.4	0.4	0.4
7c								0.4	0.4	−0.4
8p	−1	1	1	−1	−1	1	1	0.4	0.4	0.4
8c								0.4	0.4	−0.4
9p	−1	1	1	1	−1	−1	1	0.4	0.4	0.4
9c								0.4	0.4	−0.4
10p	−1	1	1	−1	−1	−1	1	0.4	0.4	0.4
10c								0.4	0.4	−0.4

Each scenario combines one of ten path signs combinations and one of four mediator residual correlation matrices, denoted by the numeric and alphabetic parts of scenario name. In the alphabetic part, 'p' = all mediator residual correlations are positive; 'a', 'b' and 'c' = the residual correlation between mediators 1&2, 1&3, and 2&3, respectively, is negative.
$\rho_{12} = \text{corr}(M^{[1]}, M^{[2]}), \rho_{13} = \text{corr}(M^{[1]}, M^{[3]}), \rho_{23} = \text{corr}(M^{[2]}, M^{[3]})$.

variations: one where path strengths vary in a symmetric manner, $\text{abs}(\boldsymbol{\alpha}) = \text{abs}(\boldsymbol{\beta}) = (0.3, 1, 1.7)$; and one where they vary in an asymmetric manner, $\text{abs}(\boldsymbol{\alpha}) = (0.3, 1, 1.7), \text{abs}(\boldsymbol{\beta}) = (1.7, 1, 0.3)$ ('abs' means absolute value) – using sign scenarios 4a, 5p and 6p as base scenarios for other parameters. We also compare uniform mediator residual correlation strengths (all with absolute value 0.4) with three other cases: low correlations (absolute value 0.1); high correlations (absolute value 0.7); and mixed correlations (absolute values including 0.1, 0.4 and 0.7) – using sign scenarios 2a, 7p and 10c as base scenarios for other parameters.

Variation in mediator-outcome confounding (exposure randomized). We examine four cases of confounding by a single variable Z: positive ($\boldsymbol{\lambda} = (1, 1, 1)^T, \delta = 1$), negative ($\boldsymbol{\lambda} = (1, 1, 1)^T, \delta = -1$), mixed but mostly positive ($\boldsymbol{\lambda} = (-1, 1, 1)^T, \delta = 1$), and mixed but mostly negative ($\boldsymbol{\lambda} = (-1, 1, 1)^T, \delta = -1$), confounding. We combine these with path coefficients and mediator residual correlations from the 22 sign scenarios, generating 88 scenarios. The distribution of Z is not of modeling interest; for data generation, we pick one that is simple to average over to calculate the true potential outcome probabilities: a discretized normal distribution with mean 0 and variance 1, with ten equal-mass points (−1.754, −1.105, −0.719, −0.411, −0.134, 0.134, 0.411, 0.719, 1.105, and 1.754).

All paths confounded. We modify the above 88 scenarios, letting the confounder influence exposure assignment: $\text{P}(X = 1|Z) = \Phi(0.5Z)$.

Small potential outcome probabilities. Based on some results from previous investigations suggesting that a small potential outcome probability might lead to bias in estimated RR-based effects, we investigate this matter specifically. We use four sign scenarios (5p, 6p, 10p, and 9p) as the basis to create four series of scenarios: in each series, scenarios share all parameters with the base scenario, but their outcome thresholds vary to let the smallest potential outcome probability range from 0.03 to 0.1 (see Table 3). The base scenarios are selected so that the series differ in which probability is the smallest (p_{00} in series 5p and 6p, and p_{01} in series 10p and 9p) and in the degree to which the four potential outcome probabilities compare to one another (being more similar in series 5p and 10p and more distant in series 6p and 9p).

In all these investigations, the ordinal mediators' thresholds are set to split them each into four cate-

Table 3
Series of small-potential-outcome-probability scenarios: ranges of potential outcome probabilities and of actual outcome prevalence

	p_{00}	p_{11}	p_{10}	p_{01}	overall outcome prevalence
Series 5p	0.03–0.10	0.14–0.31	0.07–0.19	0.07–0.19	0.08–0.21
Series 6p		0.64–0.83	0.09–0.23	0.42–0.65	0.33–0.47
Series 10p	0.07–0.19	0.07–0.19	0.14–0.31	0.03–0.10	0.07–0.19
Series 9p	0.42–0.65	0.09–0.23	0.64–0.83		0.26–0.44

gories with (approximately) equal probability mass. In all investigations except that of small potential outcome probabilities, the outcome threshold splits the outcome into two categories with equal probability mass.

For additional information, the simulations cover wide ranges in the absolute values of standardized path coefficients: 0.30 to 1.30 for α coefficients (standardized with respect to the mediator); 0.08 to 1.28 for β coefficients (standardized with respect to the mediator and outcome); and 0.18 to 0.79 for γ coefficients (standardized with respect to the outcome).

Results of simulation studies

While the investigations of variation in path/correlation signs and of variation in confounding are conceptualized sequentially (as described above), since they share the same 22 sets of values for parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\Sigma}$, it is more concise and informative to present their results together. This section therefore first reports results from (1) sign scenarios without confounding and scenarios with confounding, followed by (2) variation in path and correlation strengths, and lastly (3) small potential outcome probabilities.

Sign scenarios without confounding and scenarios with confounding. Overall, the method performs quite well for 20 out of the 22 sets of values for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\Sigma}$ (see Table 2), except the two sets defined for scenarios 6c and 9c (which will be discussed separately). Figure 2 presents results from scenarios using these 20 sets of parameters – including those with no confounding, with mediator-outcome confounding and with confounding of all paths, in the first, second and third columns. Top and bottom panels represent RD- and RR-based effects. Bias, RMSE and 95% CI coverage are presented by estimator and mediator type. The plots are violin plots, which combine density plots (the outer shell) and boxplots (the inner core). Each plot includes information from all scenarios in each category (20 sign scenarios for each plot in column 1, and 80 confounding scenarios for each plot in the other columns), with five effects per scenario (1 TE, 2 NDEs and 2 NIEs).

In these scenarios, the three estimators perform similarly. There are some differences: When estimating RD-based effects using continuous mediators, ML and WLSMV have slightly smaller bias than Bayes (although Bayes bias is also very small). When estimating RD-based effects using ordinal mediators, WLSMV is less biased than Bayes and this difference is more pronounced in the presence of confounding, but Bayes SEs are smaller, resulting in smaller RMSE. For RR-based effects, there is a slight positive bias tendency – consistent with the fact that when one quantity is divided by another that has non-zero variance, this variance induces positive bias in the magnitude of the ratio. Most of the biases are small, however. Unlike with RD-based effects, with RR-based effects, Bayes estimator performs the best with regards to both bias and variance, and this is more pronounced with ordinal mediators. Coverage is similar across sets of scenarios, estimators and mediator types, and is close to 95%. Median coverage levels are slightly under 95%.

Examination of confounding variation among these scenarios (excluding those with no confounding) shows no difference in method performance across confounding types.

Scenarios 6c and 9c are very special cases. Their parameters do not immediately suggest what might be problematic. However, if we flip the sign (or reverse the category order) of the first mediator, they are mathematically equivalent to the two setups below.

	6c equivalent		*9c equivalent*	
	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
$M^{[1]}$	1	1	1	−1
$M^{[2]}$	1	1	1	−1
$M^{[3]}$	1	1	1	−1
	$\gamma = 1$		$\gamma = 1$	
	ρ's= −.4		ρ's= −.4	

Here the mediators have the same relationships (as one another) with the exposure and the outcome, but their residuals are all negatively correlated. This results in substantial imprecision in estimating the outcome model's parameters, which entails imprecision in potential outcome probabilities computed as well as bias for probabilities far from 0.5 (due to the nonlinear transformation). Bias in estimated RDs (see Figure 3, first column, top panel) are thus substantially larger than in the non-problematic scenarios (see corresponding section of Figure 2). With RRs (Figure 3,

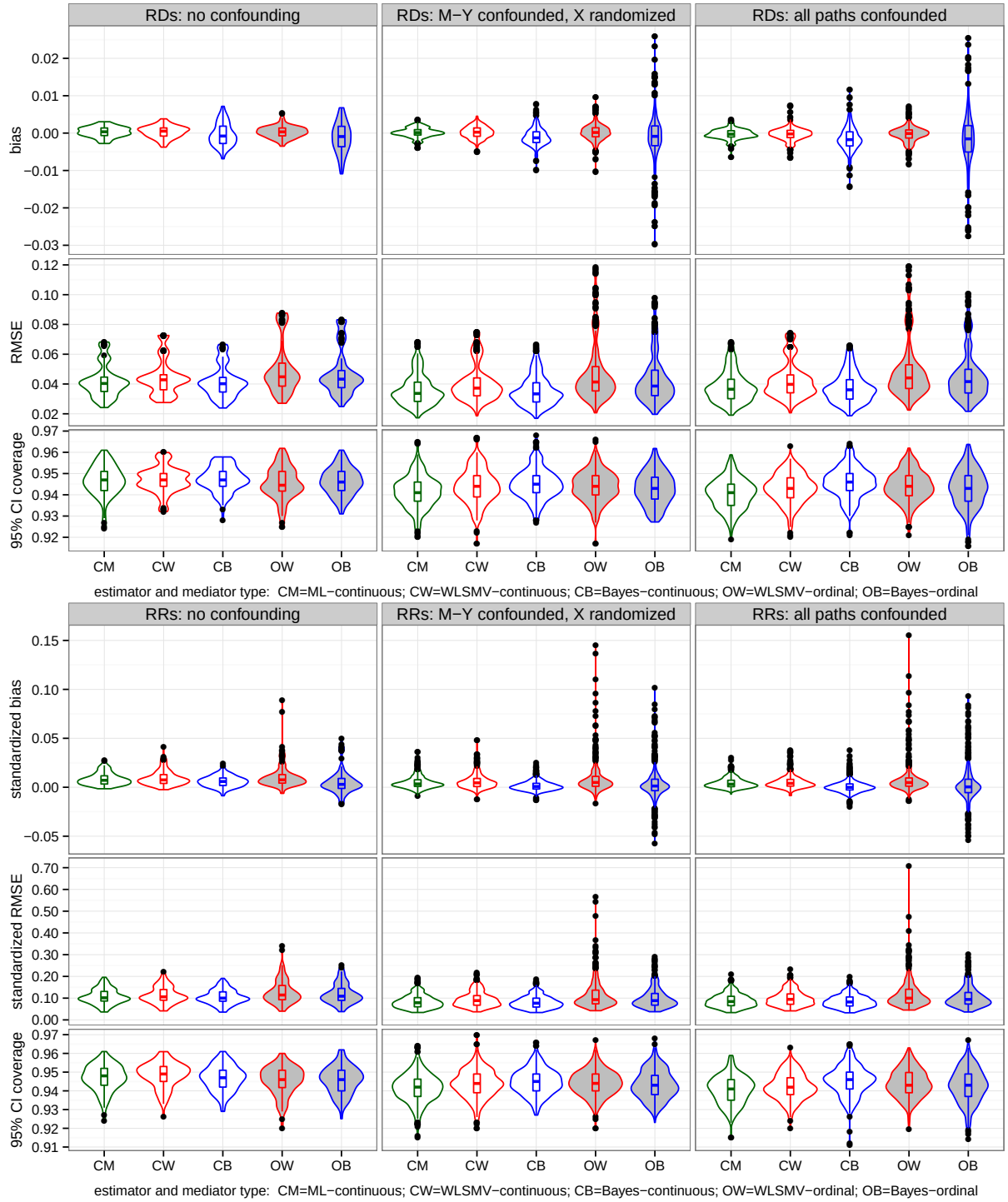

Figure 2. Results from non-problematic sign scenarios without confounding (first column), with mediator-outcome confounding (second column) and with all paths confounded (third column) – excluding problematic scenarios 6c, 9c and those based on them. Top and bottom panels represent risk difference (RD) and risk ratio (RR) based effects. Bias, root mean square error (RMSE) and 95% CI coverage are presented by estimator and mediator type. The plots are violin plots, which combine density plots (the outer shell) and boxplots (the inner core). Each plot includes information from all scenarios in each category (20 for each plot in column 1 and 80 for each plot in the other columns) with five effects per scenario (1 total, 2 natural direct and 2 natural indirect), i.e., each plot in column 1 represents 100 effects, and each plot in the other columns represents 400 effects.

first column, bottom panel), when using ordinal mediators, one effect in each of these scenarios is especially biased: $NIE(1\cdot)_{RR} = p_{11}/p_{10}$ in scenario 6c, and $NDE(\cdot 1)_{RR} = p_{11}/p_{01}$ in 9c. In both cases, the denominator is a probability that is estimated very imprecisely and with bias (SD equivalent to 65% and 77% of the true value in scenarios 6c and 9c, respectively, and bias equivalent to 12% of the true value in both – using WLSMV).

Examining scenarios 6c and 9c alongside scenarios with confounding based on them (Figure 3, second and third columns), we first note that when using continuous mediators, the ranges of bias (and RMSE, not shown) in both RD- and RR-based effects are comparable across estimators, similar to what is seen in non-problematic scenarios. When using ordinal mediators, several patterns emerge: First, for RD-based effects, WLSMV has larger bias when there is no confounding, but Bayes has larger bias when there is confounding. Second, with ordinal mediators, the combination of 6c and 9c each with one of the four mediator-outcome confounding setups results in greater bias in the most biased RR-based effect, but combinations with the other three confounding setups substantially reduce the bias. (This is why the second and third columns in Figure 3 have the same number of points representing large biases in RR-based effects as the first column, and not four times as many points.) The key here is variation in the potential outcome probability that is the denominator of the effect. In confounding scenarios with reduced bias, this probability is larger, and closer to 0.5, than in the base scenario (6c/9c), and is thus estimated with less bias. In confounding scenarios with increased bias, this probability is the same as in the base scenario, but confounding results in larger variance, which induces larger bias in the ratio. Third, with ordinal mediators, Bayes has better RMSE (not shown) than WLSMV for both RD- and RR-based effects, similar to what is seen in non-problematic scenarios. With respect to coverage, in confounding scenarios, coverage seems largely comparable across estimators and mediator types, albeit slightly lower with ML estimator. In no-confounding scenarios, coverage is noticeably worse when using ML or WLSMV. This may be due to the different methods by which CIs are obtained. When using Bayes estimator across the board and when using ML or WLSMV for the confounding case, we rely on some sampling method to obtain CIs – resampling from the data with ML and WLSMV, and sampling the posterior distribution with Bayes estimator. ML and WLSMV CIs for the non-confounding case, however, are Delta method-based and may be more prone to low coverage in these problematic scenarios.

Path and correlation strengths. The method performs well across the variations in magnitudes of path coefficients and mediator residual correlations, with bias, MSE and coverage comparable to those in the base scenarios (see more in Web Appendix B).

RR-based effects with small potential outcome probabilities. Figure 4 presents bias in RR-based effects in the four series of small-potential-outcome-probability scenarios. Where bias is present, the smaller the probability, the larger the bias. Bias is not uniform, however, over the different series or over the different effects. Comparing the series, overall bias is related to the rareness of the actual outcome: Series 9p has the least bias, because even though it has small p_{01}, this is a truly counterfactual probability which does not represent actual outcome prevalence; actual outcome prevalence in the two exposure conditions is better indicated by p_{11} and p_{00} which in this series are not quite small. The other three series, which have more bias, either have low outcome prevalence in one exposure condition (series 5p and 6p, outcome prevalence ranging from 0.03 to 0.10 in exposure condition 0), or relatively low outcome prevalence in both exposure conditions (series 10p, outcome prevalence in both conditions ranging from 0.07 to 0.19).

Within a series or a scenario, which specific RR-based effects are biased depends on which potential outcome probabilities are estimated with bias, and which probabilities that serve as denominators of effects have large variance. Here it is relative bias and relative variance (i.e., raw bias and raw variance divided by the probability) that matter. Several factors affect these two elements: (i) truly counterfactual probabilities (p_{10}, p_{01}) tend to be estimated more imprecisely than partially observed probabilities (p_{00}, p_{11}); (ii) probabilities farther away from 0.5 tend to have larger raw bias and smaller raw variance; (iii) the larger the uncertainty in model parameter estimates, the larger the variance and bias of predicted potential outcome probabilities; and (iv) relative variance and relative bias tend to be larger for smaller probabilities. In the series of scenarios considered, the largest biases are seen in effects whose denominators are the smallest potential outcome probability (TE_{RR} and $NDE(\cdot 0)_{RR}$ in series 5p and 6p, and $NDE(\cdot 1)_{RR}$ in series 10p and 9p). This seems to be a general trend, but not a rule, because the factors listed above may work together in complicated ways and may result in another effect being the most biased; an example is $NIE(1\cdot)_{RR}$ in scenario 6c, whose denominator is $p_{10} = 0.22$, which is not small and not the smallest potential outcome probability ($p_{00} = 0.06$), but is estimated with by far the largest relative variance and relative bias.

Comparing estimators, in most of the effect-by-series plots in Figure 4, and in all those with the most bias, Bayes estimator is the least biased, bringing bias close to zero even for the lowest end of the outcome proba-

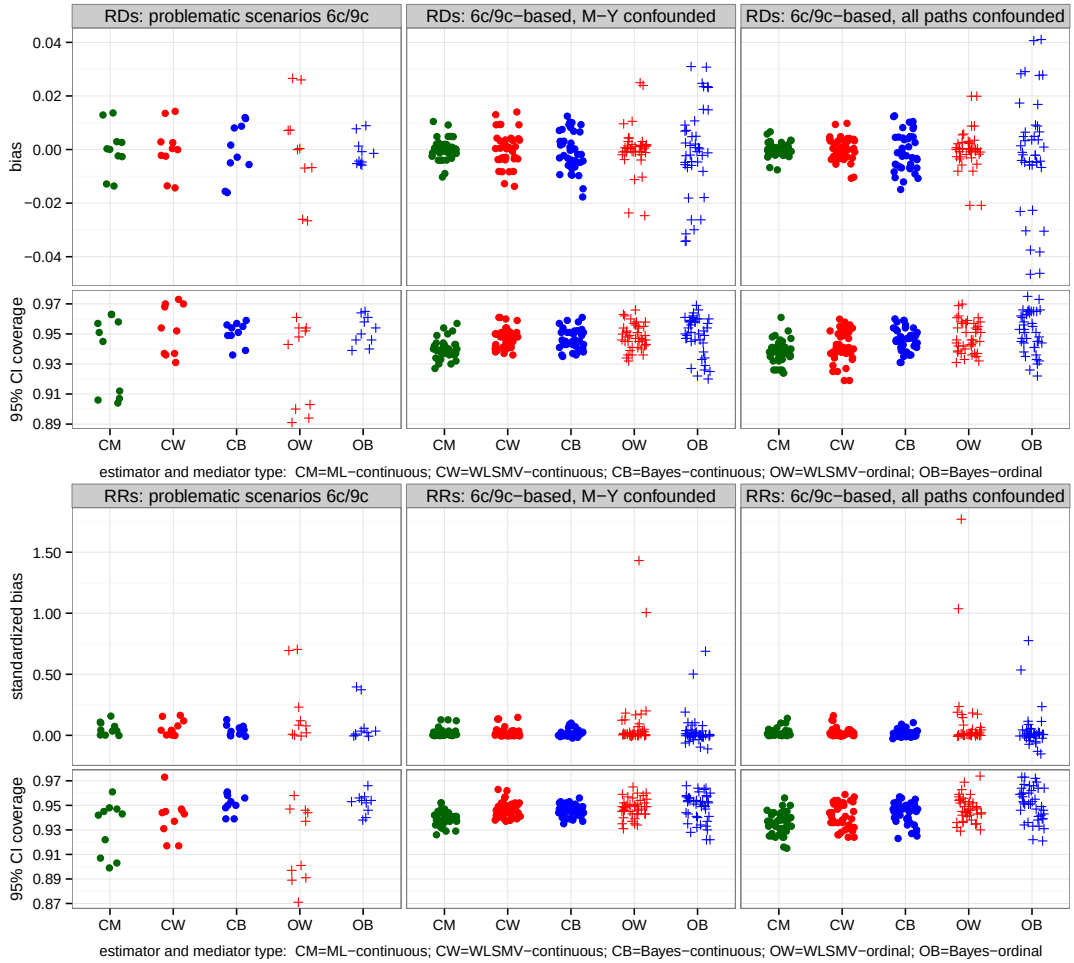

Figure 3. Results from problematic sign scenarios 6c and 9c (first column), and scenarios based on them that also have mediator-outcome confounding (second column) and all paths confounded (third column). Top and bottom panels represent risk difference (RD) and risk ratio (RR) based effects. Bias and 95% CI coverage are presented by estimator and mediator type, using dot plots. Each plot includes information from all scenarios in each category (2 in column 1, and 8 in the other columns) with five effects per scenario (1 total, 2 natural direct and 2 natural indirect), i.e., each plot in column 1 represents 10 effects, and each plot in the other columns represents 40 effects.

bility range. This is consistent with prior results favoring Bayes for RR-based effects with ordinal mediators. In these small-potential- outcome-probability situations, Bayes is also favored when using continuous mediators.

Summary of simulation results. With the commonly used ML and WLSMV estimators in Mplus, the method performs well in most cases, for RD-based effects generally and for RR-based effects when using continuous mediators. When using ordinal mediators and estimating RR-based effects, Bayes performs better than WLSMV. In the special case where the mediators have the same relationships, as one another, with the exposure and the outcome, but their residuals are negatively correlated (and in mathematically equivalent situations), there is great uncertainty in parameter estimates and the effects have larger bias. In this case, with continuous mediators, the three estimators perform similarly for both RD- and RR-based effects, with the exception that Bayes estimator has smaller RMSE for RR-based effects; with ordinal mediators, WLSMV is better for RD-based effects and Bayes for RR-based effects. When one or more potential outcome probabilities are small and RR-based effects are of interest, Bayes is generally the best estimator, regardless of mediator type. Additional simulations show that in the extreme case with a semi-positive-definite mediator residual correlation matrix, Bayes is also the best: ML fails completely, WLSMV either fails to fit the model or greatly overestimates variance, while Bayes successfully fits the model, overestimates variance but to a much lesser ex-

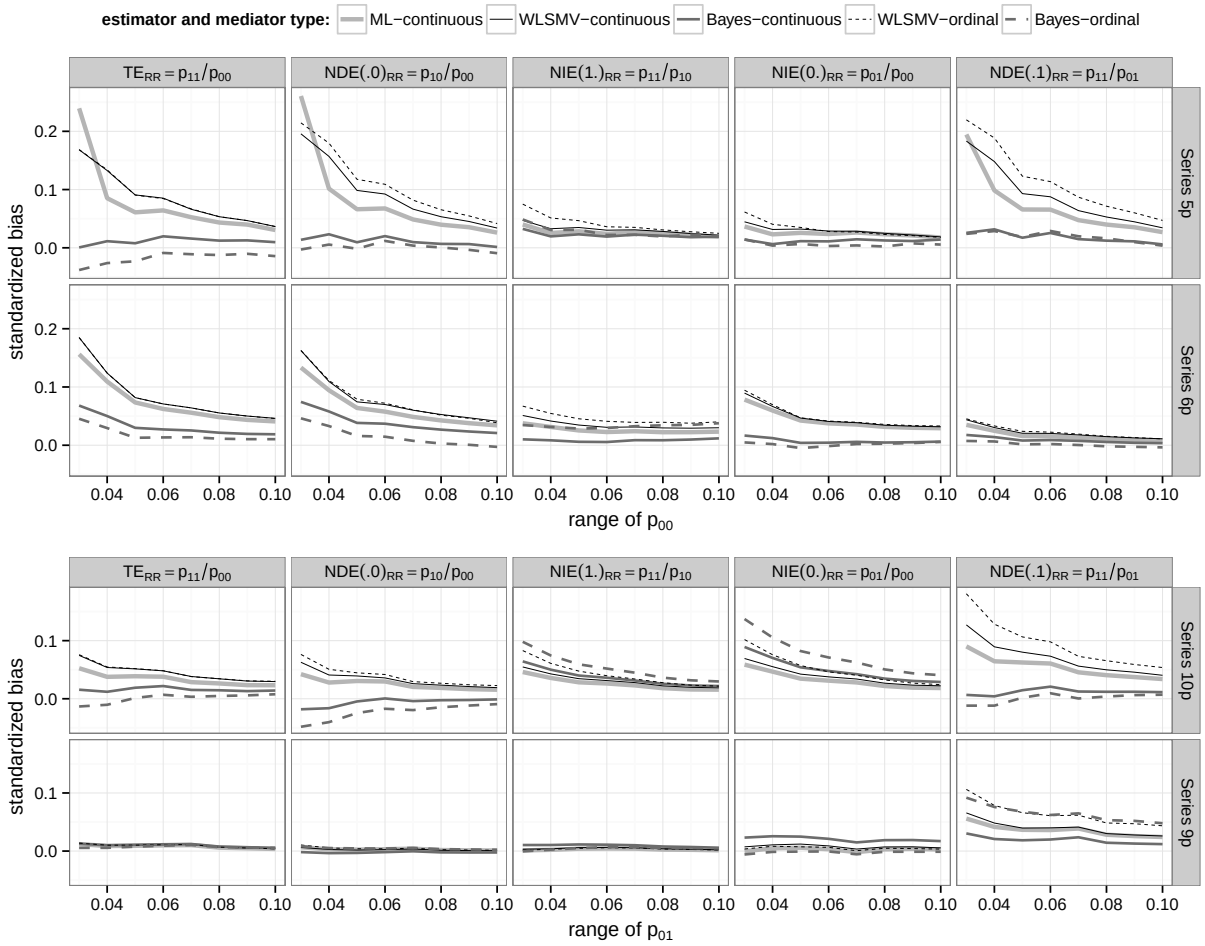

Figure 4. Bias of RR-based effects in four small-potential-outcome-probability series of scenarios, by estimator and mediator type. Biases are standardized by dividing by the true effects.

tent, and has substantially smaller RMSE (see more in Web Appendix C).

Application to the PAS trial

We apply the method to data from the PAS trial, comparing the parent-and-student combined intervention to the control condition of regular curriculum. The mediators examined are adolescent self-control, adolescent-reported parental rules about alcohol, parent-reported rules, adolescent attitudes about alcohol and parent attitudes about adolescent drinking, at 10-month follow-up. The outcome is weekly drinking at 22-month follow-up. All mediator/outcome models control for baseline age, gender, education track (vocational or academic), parent education (the higher attainment of parents, if two parents), and religion (no religion, Christianity, Islam or other religions). To strengthen the plausibility of the no unmeasured confounding assumptions, each mediator model controls for the mediator's baseline measure; and the outcome (as well as the mediator models) control for baseline drinking frequency. In addition, each of the adolescent-reported mediators is allowed to be predicted by baseline measures of the others; each parent-reported mediator is allowed to be predicted by the other's baseline measure; and adolescent-reported rules is allowed to be predicted by baseline parent-reported rules and parent attitudes. For detailed description of the measures, see Koning et al. (2010). For mediators that are highly skewed (adolescent- and parent-reported rules, and adolescent and parent attitudes), controlling for baseline measures also helps bring their residuals closer to normally distributed, which is important because the method relies on combining normal error terms.

For this illustrative example, to keep things simple, we restrict the sample to adolescents with fully observed baseline data (n=1178, including 536 students in 36

classes in five intervention schools, and 642 students in 49 classes in four control schools); all inferences made here are with respect to this specific sample. With missing data on the outcome (17.0%) and mediators (3.5% for student-, and 18.4% for parent-reported mediators, on average), we use full-information maximum likelihood estimation, assuming missing at random given observed data. Maximum likelihood estimation is also appropriate with the continuous mediators. The analysis incorporates clustering of students in classes (see Mplus input in Web Appendix E). We obtain 95% CIs of potential outcome probabilities and causal mediation effects via bootstrapping, with 500 bootstrap samples.

In the model with five mediators, the intervention positively predicts all mediators, and four mediators (except parent-reported rules) negatively predict the outcome. Parent-reported rules is strongly correlated with both adolescent-reported rules and parent attitudes, and thus is non-significant after these other two variables are accounted for. This variable is removed, and the model with four mediators (see Figure 5) is the final model. This model is consistent with previous findings – three of the four mediators we retain (adolescent self-control, adolescent-reported rules, and parent attitudes) were found to mediate the intervention's effect on weekly drinking onset (Koning et al., 2010). As all mediator residual covariances are positive and the predicted potential outcome probabilities (see next paragraph) are not small, this is not a problematic or special case.

The potential prevalence of weekly drinking at 22 months is estimated to be p_{00}=31.4% (CI=27.8,35.5%) had the whole sample been in the control condition; p_{11}=18.1% (CI=14.3,21.7%) had the whole sample been in the intervention condition; and p_{10}=22.0% (CI=17.3,26.3%) had the whole sample participated in the intervention but the mediators been kept at control levels. For illustrative purposes, here we include effects on both RD and RR scales; usually only one is needed.

$$
\begin{aligned}
TE_{RD} &= -13.3\% \quad (95\%CI = -18.9, -7.9\%)\\
NDE(\cdot 0)_{RD} &= -9.4\% \quad (95\%CI = -15.0, -3.0\%)\\
NIE(1\cdot)_{RD} &= -3.9\% \quad (95\%CI = -5.5, -2.5\%)\\
TE_{RR} &= 0.58 \quad (95\%CI = 0.45, 0.72)\\
NDE(\cdot 0)_{RR} &= 0.70 \quad (95\%CI = 0.55, 0.87)\\
NIE(1\cdot)_{RR} &= 0.82 \quad (95\%CI = 0.76, 0.88)
\end{aligned}
$$

If the model is correctly specified and the identifying assumptions hold (this should be judged carefully!), we could interpret these effects as causal: If the whole sample had been in the intervention (as opposed to control condition), this would have lowered weekly drinking prevalence at 22 months by 13.3 (CI=7.9,18.9) percentage points or by a ratio of 0.58 (CI=0.45,0.72) (the total effect). If the whole sample had participated in the intervention but the mediators (adolescent self-control, adolescent-reported parental rules, adolescent attitudes and parent attitudes) had been kept at control levels and not allowed to change, this hypothetical condition would also have lowered weekly drinking prevalence, relative to the control condition, by 9.4 (CI=3.0,15.0) percentage points or by a ratio of 0.70 (CI=0.55,0.87) (the direct/unmediated effect). Relative to this hypothetical condition, a normal intervention condition (where the whole sample had participated in, and their mediators been free to change as a result of, the intervention) would have weekly drinking prevalence that is lower by 3.9 (CI=2.5,5.5) percentage points or by a ratio of 0.82 (CI=0.76,0.88) (the indirect/mediated effect).

Researchers familiar with traditional mediation methods may have noted that an analogous traditional analysis would involve fitting the same structural equation model and computing the *total indirect effect* – the sum of four mediator-specific indirect effects, each a product of exposure-to-mediator and mediator-to-outcome coefficients. The estimates of this indirect effect – unstandardized, and standardized with respect to the outcome – are −0.17 (standard error [SE] = 0.04) and −0.13 (SE=0.03), respectively. The corresponding direct effect estimates are −0.35 (SE=0.11) and −0.26 (SE=0.09) – see additional model outputs in Figure 5. This result is interpreted in association terms: holding confounders constant, on average students in the intervention arm have lower 'outcome' than those in the control arm, and this difference consists of a part mediated by the mediators and an unmediated part, of magnitudes −0.13 (SE=0.03) and −0.26 (SE=0.09) standard deviations, respectively. The 'outcome' here is not the binary outcome variable of interest (weekly drinking), but the latent continuous variable underlying this binary variable.

By defining effects based on the potential outcome framework, causal inference methods make it possible to go beyond association to make inference about causality, contingent on assumptions. With the proposed method, the conversion of parameter estimates to potential outcome probabilities makes it possible to discuss effects in terms of the outcome variable of interest rather than a latent underlying variable.

Discussion

This paper presents an approach for estimating mediation effects when there are multiple continuous or ordinal mediators and the outcome is binary, which involves (i) fitting probit/normal models for the mediators and the outcome, (ii) using parameter estimates and confounder data to predict potential outcome probabilities, and (iii) using such probabilities to compute TE, NIE

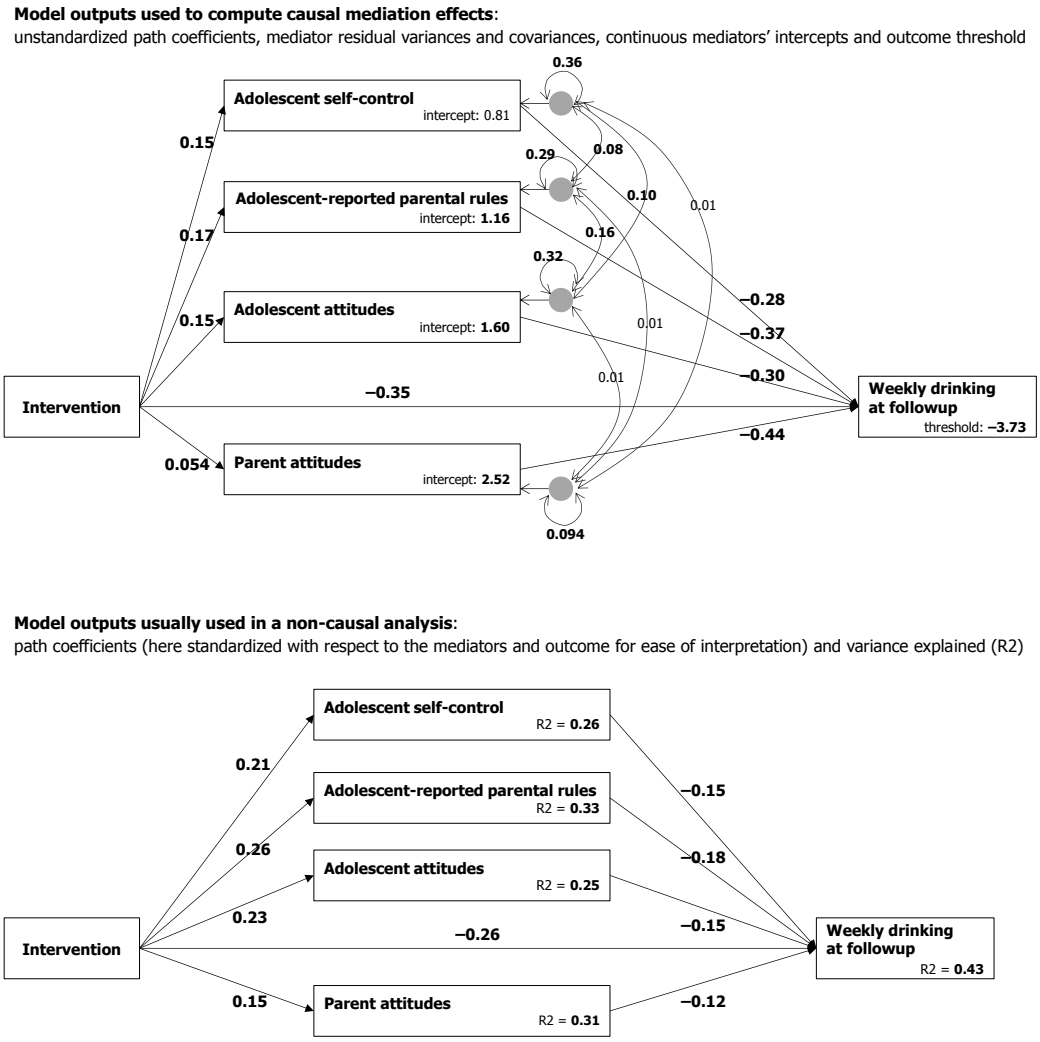

Figure 5. The final model for the illustrative example, which controls for age, gender, education track, parent education, religion, baseline measures of mediators and baseline drinking frequency (see text for detail). The top diagram presents model outputs used in computing causal mediation effects. These are contrasted with the bottom diagram including outputs usually used in non-causal analysis. Standard errors and p-values are not included to avoid cluttering. Statistically significant quantities (with p-value<0.05) are bolded.

and NDE. The proposed method performs well with all the estimators considered in most simulated scenarios, with special cases where one estimator performs better. Applying to a real data example, we partitioned the TE of an adolescent alcohol prevention intervention into a NIE and a NDE, both expressed in terms of reduction in drinking prevalence; such findings may be useful to prevention scientists as well as program managers and policy makers. This method is relevant to analyses of social/behavioral interventions designed to impact multiple mediating pathways, as well as naturally occurring exposures that act on an outcome through multiple mechanisms.

We offer several suggestions for the use of this method, assuming that in most cases causal mediation analysis incorporates confounders. The choice of an estimator (and in some cases the effects to estimate) is simple, based on the following rules-of-thumb:

- In general, with continuous mediators, all three estimators can be used. ML and WLSMV are slightly favored for RDs. Bayes is slightly favored for RRs.

- In general, with ordinal/mixed mediators, use WLSMV for RDs and Bayes for RRs.
- A special case: If one or more potential outcome probabilities are expected or found to be small, and RRs are of interest, use Bayes, regardless of mediator type.
- A problematic but uncommon case: If the mediator residual covariance matrix is semi-positive-definite, use Bayes and estimate RDs (not RRs).
- A problematic but probably rare case: With ordinal/mixed mediators, (i) if the mediators have similar relationships (to one another) with the exposure and the outcome, but their residuals are all negatively correlated (you can check this after fitting the model), and (ii) in mathematically equivalent situations (i.e., you get (i) after changing the sign(s) of certain mediators(s) and/or of the outcome), there are options to minimize bias: (1) estimate RDs instead of RRs and use WLSMV to do so; or if RRs are desired (2) estimate RRs using Bayes (less biased than WLSMV), or (3) choose a TE decomposition less affected by bias. The third option may be hard, but if the case is similar to scenario 6c, use $TE_{RR} = NIE(0\cdot)_{RR} \times NDE(\cdot 1)_{RR}$; if it is similar to 9c, use $TE_{RR} = NDE(\cdot 0)_{RR} \times NIE(1\cdot)_{RR}$.

In any causal inference analysis, we urge the researcher to thoroughly consider the method's assumptions given the data to be analyzed, to avoid making faulty causal claims. With the proposed method we have discussed identification assumptions, with a focus on different types of confounding. The model itself is a combination of assumptions about how the variables influence one another and about the distributions of the errors; for this method an important assumption is normally distributed errors. Since many of the measures used in social research are non-normal, it is important when using this method to check whether residuals are close to normally distributed.

Our next step in the development of this method will be to evaluate its performance under violation of some model assumptions (including non-normal residuals, and true outcome models different from probit), and in a broader range of situations (e.g., an exposure-mediator interaction, which is not covered by simulations in this paper, or ordinal mediators with heavier mass at one end). This method should also be compared in future work to other methods for similar purposes, including VanderWeele and Vansteelandt's (2013). The strategy of using a multivariate distribution to allow mediator residual dependence could potentially be extended to accommodate other mediator types. Another area for future work is sensitivity analysis for unmeasured confounding, especially confounding of the mediator-outcome relationships, in this multiple-mediator setting.

While these simulations consider causally-unordered mediators (i.e., one mediator does not cause another), the proposed method can be used with causally-ordered mediators when the purpose is to estimate their combined mediation effect. In this case, the assumption of no exposure-induced mediator-outcome confounding applies to the collection of mediators as a whole, i.e., there is no variable that is influenced by the exposure and that influences the outcome and one or more mediators in the collection, but it is fine for mediators within the collection to influence one another. If this and the other assumptions hold, the method correctly estimates potential outcome probabilities.

If the interest is in path-specific effects in the context of causally-ordered mediators, readers are referred to other work, e.g., Daniel et al. (2015); Imai and Yamamoto (2013); VanderWeele and Vansteelandt (2013); Vanderweele, Vansteelandt, and Robins (2014).

While this study considers one outcome, models with multiple outcomes are common in SEM research. When such models are used for causal analysis, it is important to consider the component for each outcome: which mediators are relevant, which confounders are needed, and how likely the identification assumptions are to hold. A model combining multiple outcomes may include more confounder and/or mediator variables than needed for each of the outcomes, and it is unclear how such combined analysis compares to separate analyses; this needs to be investigated in future work.

This study investigates the estimation of RD- and RR-, but not OR-based effects. We postulate that the proposed method would estimate ORs less well than it estimates RDs and RRs. Four probabilities need to be estimated for one OR; any bias in any of the four probabilities and variance in the two denominator probabilities contribute to biasing the OR. (The RR, in contrast, involves only two probabilities, with only one being in the denominator.) In cases with potential outcome probabilities close to 0.5, however, the method may still work. Also, the Bayes estimator, which works well with RRs, might have acceptable performance with ORs. These are questions requiring future investigation.

The simulation studies consider independent units, but the alcohol prevention study involves clustered data. To keep things simple for an illustrative example, we estimate marginal means, variances and covariances, correcting standard errors for clustering. However, the model does not allow parameters to vary across clusters. Future work should examine how to generalize this method to a multilevel setting so that potential outcome probabilities could be expressed as a function of individual factors and cluster membership.

To facilitate application, Web Appendix D includes generic Mplus inputs covering a mix of two continuous mediators, two ordinal mediators and two confounders, which can easily be adapted to other situations. These are accompanied by R code that performs bootstrapping when using WLSMV/ML, or processes the posterior distribution when using Bayes. The Mplus input for the final model of the illustrative example is also included (Web Appendix E).

References

Albert, J. M. (2012). Distribution-free mediation analysis for nonlinear models with confounding. *Epidemiology*, *23*(6), 879–88.

Ananth, C. V., & VanderWeele, T. J. (2011). Placental abruption and perinatal mortality with preterm delivery as a mediator: Disentangling direct and indirect effects. *American Journal of Epidemiology*, *174*(1), 99–108.

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. *Unpublished technical report*. Retrieved from http://statmodel2.com/download/Bayes3.pdf

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Bennett, A. C., Rankin, K. M., & Rosenberg, D. (2012). Does a medical home mediate racial disparities in unmet healthcare needs among children with special healthcare needs? *Maternal and Child Health Journal*, *16 Suppl 2*, 330–8.

Coffman, D. L., & Zhong, W. (2012). Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods*, *17*(4), 642–64.

Daniel, R., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*, 1–14.

Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine*, *23*(17), 2713–28.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–34.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, *21*(2), 141–171.

Kelly, J. F., Hoeppner, B., Stout, R. L., & Pagano, M. (2012). Determining the relative importance of the mechanisms of behavior change within Alcoholics Anonymous: A multiple mediator analysis. *Addiction*, *107*(2), 289–99.

Koning, I. M., Van Den Eijnden, R. J., Verdurmen, J. E., Engels, R. C., & Vollebergh, W. a. (2011). Long-term effects of a parent and student intervention on alcohol use in adolescents: A cluster randomized controlled trial. *American Journal of Preventive Medicine*, *40*, 541–547.

Koning, I. M., van den Eijnden, R. J. J. M., Engels, R. C. M. E., Verdurmen, J. E. E., & Vollebergh, W. a. M. (2010). Why target early adolescents and parents in alcohol prevention? The mediating effects of self-control, rules and attitudes about alcohol use. *Addiction*, *106*(3), 538–46.

Koning, I. M., Vollebergh, W. a. M., Smit, F., Verdurmen, J. E. E., Van Den Eijnden, R. J. J. M., Ter Bogt, T. F. M., ... Engels, R. C. M. E. (2009). Preventing heavy alcohol use in adolescents (PAS): Cluster randomized trial of a parent and student intervention offered separately and simultaneously. *Addiction*, *104*(10), 1669–78.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York, NY: Taylor & Francis.

Muthén, B. O. (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. *Unpublished working paper*. Retrieved from http://www.statmodel2.com/download/causalmediation.pdf

Muthén, B. O., & Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 12–23.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modelling with categorical and continuous outcomes. *Unpublished technical report*. Retrieved from http://www.statmodel.com/bmuthen/articles/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

Nandi, A., Glymour, M. M., Kawachi, I., & VanderWeele, T. J. (2012). Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart dis-

ease, diabetes, and stroke. *Epidemiology*, *23*(2), 223–32.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty and artificial intelligence* (pp. 411–420). San Francisco: Morgan Kaufmann.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.

Petersen, M. L., Sinisi, S. E., & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, *17*(3), 276–84.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (Third ed.). Phildadelphia, PA: Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes*. *Scandinavian Journal of Statistics*, *31*, 161–170.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 149–160.

Schuck, A. M., & Spatz, C. (2001). Childhood victimization and alcohol symptoms in females: causal inferences and hypothesized mediators. *Child Abuse & Neglect*, *25*, 1069–1092.

Smith, P. M., Smith, B. T., Mustard, C. a., Lu, H., & Glazier, R. H. (2013). Estimating the direct and indirect pathways between education and diabetes incidence among Canadian men and women: a mediation analysis. *Annals of Epidemiology*, *23*(3), 143–9.

Subbaraman, M. S., Lendle, S., van der Laan, M. J., Kaskutas, L. A., & Ahern, J. (2013). Cravings as a mediator and moderator of drinking outcomes in the COMBINE study. *Addiction*, *108*(10), 1737–44.

Ten Have, T. R., & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research*, *21*(1), 77–107.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, *2*, 457–468.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*(12), 1339–1348.

VanderWeele, T. J., & Vansteelandt, S. (2013). Mediation Analysis with Multiple Mediators. *Epidemiologic Methods*, *2*(1), 95–115.

Vanderweele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, *25*(2), 300–6.