Lower-level mediation with binary measures

Article *in* Statistics and its interface · January 2019

DOI: 10.4310/SII.2019.v12.n4.a2

CITATION
1

READS
17

3 authors:

Haeike Josephy
Ghent University
12 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Tom Loeys
Ghent University
116 PUBLICATIONS **1,331** CITATIONS
SEE PROFILE

Sara Kindt
Ghent University
9 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Statistics and Its Interface Volume 12 (2019) 511–526

Lower-level mediation with binary measures

Haeike Josephy*, Tom Loeys, and Sara Kindt

In recent literature, researchers have put a lot of time and effort in expanding mediation to multilevel settings. Unfortunately, such extensions are often limited to continuous settings, whereas research on multilevel mediation with binary mediators and outcomes remains rather sparse. Additionally, in lower-level mediation, the effect of the lower-level mediator on the outcome may oftentimes be confounded by an (un)measured upper-level variable. When such confounding is left unaddressed, the effect of the mediator, and hence the causal mediation effects themselves, will be estimated with bias. In linear settings, bias due to unmeasured additive upper-level confounding is often remedied by separating the effect of the mediator into a within- and between-cluster component. However, this solution is no longer valid when considering binary outcome measures. To assess the severity of this transgression, we aim to tackle lower-level mediation in binary settings from a counterfactual point of view, with a special focus on small clusters. We do this by 1) providing non-parametrical identification assumptions of the direct and indirect effect, 2) parametrically identifying these effects based on appropriate modelling equations, 3) considering estimation models for the mediator and the outcome, and 4) estimating the causal effects through an imputation algorithm that samples counterfactuals. Since steps three and four can be completed in various ways, we compare the performance of three different estimation models (an uncentered and centred separate modelling method, and a joint approach), and two different ways of predicting random effects (marginally versus conditionally). Employing simulations, we observe that the joint approach combined with a marginal generation of random effects performs best when sample sizes are sufficiently large. Additionally, we illustrate our findings with data from a crossover study that assesses the impact of experimentally induced goal conflict on the helping behaviour of partners of individuals with chronic pain.

Keywords and phrases: Multilevel mediation, Binary measures, Unmeasured confounders, Counterfactuals.

1. INTRODUCTION

We must acknowledge that clustered or multilevel data have become protagonists in numerous research fields. In this type of data, we always encounter a specific kind of hierarchy, where usually, two levels can be distinguished: lower-level measurements are nested within clusters or upper-level units. Examples of such hierarchically nested entities consist of relatives nested within a family, students within classrooms, or measurement moments within an individual. These lower-level measures show dependencies amongst each other, as measures arising from within a family, a classroom, or an individual, will be more alike than data arising from two random units. Analyses that either ignore these dependencies or inappropriately aggregate the data across levels, will often lead to invalid inferences (Snijders and Bosker, 1999; Raudenbush and Bryk, 2002). Over the course of decades, two major frameworks have been put forward that are able to deal with such correlations: Mixed-effect Models (MM) and Structural Equation Models (SEM). Although SEM holds several advantages over its MM counterpart, both frameworks turn out to be entirely equivalent when considering balanced multilevel data within a random intercept model (Rovine and Molenaar, 2000; Curran, 2003; Bauer, 2003).

Taking the usefulness of multilevel designs into account, expanding mediation analysis to hierarchical settings has become an increasingly popular topic (Bauer et al., 2006; VanderWeele and Vansteelandt, 2009; Zhang et al., 2009; Preacher et al., 2010; Preacher, 2015; Tofighi and Kelley, 2016). When looking at the effect of a randomised exposure that varies within clusters, researchers often consider a design where the mediator and outcome too, are measured at the lower-level; this type of mediation is appropriately termed lower-level mediation. Until recently, the literature on lower-level mediation has almost exclusively relied on extending the product-of-coefficients approach to multilevel settings (Judd et al., 2001; Kenny et al., 2003; Bauer et al., 2006; Preacher et al., 2010). Unfortunately, this procedure does not offer a general definition of the causal effects that is applicable beyond the few (linear) statistical models considered. Furthermore, these extensions have mostly been executed without due attention to the interpretation of the effects as causal parameters, nor to the underlying assumptions needed to identify these. Some researchers have tried to surmount these shortcomings by tackling multilevel mediation from a counterfactual perspective (Imai et al., 2010a; VanderWeele, 2010b,a; Josephy et al., 2015). This has proven very fruitful, as this framework is able to explicate the assumptions underlying multilevel mediation, put forward a general (non-parametrical) definition of the causal effects, as well as

*Corresponding author.

identify these effects based on appropriate (parametrical) models (Pearl, 2001; VanderWeele and Vansteelandt, 2009; Imai et al., 2010a; VanderWeele, 2010b; Pearl, 2012).

1.1 Estimation of the causal mediation effects in four steps

When relying on the counterfactual framework to unbiasedly estimate the causal mediation effects, we need to consider four consecutive steps.

In a **first step**, we define non-parametrical expressions for the direct and indirect effect. For a continuous outcome, this is usually achieved on a linear scale (VanderWeele, 2010b; Josephy et al., 2015), while a linear-, risk ratio- (RR), and odds ratio (OR)-type definition have been used for categorical measures (Imai et al., 2010b; VanderWeele, 2013; Loeys et al., 2013; Bind et al., 2016). We will focus on deriving these expressions on a linear scale.

As a second part of this step, the assumptions needed to identify the above-mentioned effects are recited. One very important assumption in lower-level mediation studies entails the absence of unmeasured upper-level confounders of the mediator-outcome relationship. This type of confounding (also labeled upper-level endogeneity, Wooldridge (2010)), is very common in many contexts and may lead to serious bias in the estimation of the intervening effect. Hence, we will place emphasis on this assumption in particular.

In a **second step**, we identify parametrical expressions for the causal mediation effects based on modelling equations that satisfy the assumptions explicated in the previous step. Traditionally, most such attempts were made with a continuous scaled mediator and outcome in mind (VanderWeele, 2010b; Josephy et al., 2015).

In the **third step**, we require unbiased and efficient estimation of the regression coefficients of the mediator and outcome models. This unbiasedness depends upon the assumptions mentioned during the first step of this process; if, for example, the assumption of 'no upper-level endogeneity of the mediator-outcome relation' is not met, a traditional multilevel model for the outcome (with the mediator as a predictor) will estimate its regression coefficients with bias (Zhang et al., 2009; Josephy et al., 2015). In two-level linear settings, Centering Within-Clusters (CWC), was proposed to solve such potential confounding issues (Neuhaus and Kalbfleisch, 1998). Unfortunately, when the outcome is binary, CWC no longer yields proper parameter estimates, although the resulting bias may often be small (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010).

Alternatively, the mediator and outcome can also be modelled jointly under a slightly more stringent set of conditions (Bauer et al., 2006; Josephy et al., 2015). Such a joint modelling approach allows for unmeasured cluster-specific common causes of the mediator and the outcome, by estimating a covariance term between both random intercepts (Bauer et al., 2006; Skrondal and Rabe-Hesketh, 2014). Consequently, we aim to focus on joint modelling in order to confront issues with upper-level endogeneity of the mediator-outcome relation.

Finally, a **fourth step** aims to estimate the causal mediation effects themselves. Typically, the expressions derived during the second step are conditional on the cluster-specific random effects. In linear settings, these effects cancel each other out when the expressions are defined on a difference-scale, but unfortunately, this is no longer the case with a binary outcome. Consequently, if we want to obtain expressions for the mediation effects that are marginalised over the random effects, we need to sample these effects from their assumed distribution and average them out. Researchers can rely on two possible mechanisms for the sampling of random effects: a first possibility draws the random effect from a marginal zero-centred distribution (i.e., the marginal sampling (co)variances), while a second relies on a distribution that is conditional on the cluster identifier (i.e., the posterior (co)variances) (Tingley et al., 2014). Since it has been stated that the latter might underestimate the variance of the random effects distribution (Skrondal and Rabe-Hesketh, 2004), we aim to quantify and compare the performance of both methods.

1.2 Our work

In summary, we aim to investigate which multilevel estimation models are able to effectively eliminate unmeasured upper-level confounding of a binary mediator and outcome. In addition, we will focus on a randomised binary exposure that varies within small clusters, as limited group sizes have proven difficult for the available estimation techniques (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995). These settings are often encountered in practice, e.g. when studying dyads (McMahon et al., 2003), twins (Ortqvist et al., 2009), or few repeated measures within each individual (Senn, 2002). On top of this, we want to evaluate if, and how, the link-function and/or a conditional versus marginal sampling of the upper-level residuals, may affect the estimation of the mediation effects.

To answer these questions, we conduct a large simulation study in which we compare three estimation models for mediator and outcome (an uncentered separate modelling approach, a separate approach that relies on CWC, and a joint method), two link-functions for our binary measures (*logit* and *probit*), and two ways in which to generate the random effects (marginally vs. conditionally). In this respect, our work distinguishes itself from other papers on lower-lever mediation, as most of these either 1) do not offer a generalisable approach to lower-level mediation from a counterfactual point-of-view, 2) do not investigate the case of a binary outcome and/or mediator, or 3) do not evaluate performance measures based on extensive simulation

studies (Judd et al., 2001; Kenny et al., 2003; Bauer et al., 2006; Raykov and Mels, 2007; Montoya and Hayes, 2017; Vuorre and Bolger, 2017).

2. ILLUSTRATING EXAMPLE

We consider data from a crossover study that aims to assess the impact of experimentally induced goal conflict (i.e., the amount of interference between helping your partner and other goals) on the helping behaviour of partners of individuals with chronic pain (ICP) (Kindt et al., 2018). During this study, 68 couples[1] (with one person having chronic pain) were asked to perform a series of household activities, while the sequence of goal conflict was randomly manipulated: partners were either first asked to stay available for help while simultaneously working on a puzzle task (i.e., the goal conflict condition) and then simply asked to be available (i.e., the control condition without goal conflict), or the other way round. After each series of chores, couples reported on several intra- and interpersonal outcomes, as well as the partners' quantity and quality of help. We will focus on the effect of goal conflict (a binary exposure) on the amount of help provided by the ICP's partner (the binary outcome, high vs. low amount of help). Additionally, we wanted to check if this relation is mediated by the partner's amount of autonomous helping motivation, as perceived by the ICP (a binary mediator, high vs. low amount of helping motivation). As all three dichotomous variables are measured within every couple and each couple is only measured at two time points, we and up with a binary lower-level mediation question and a limited cluster size.

3. STEP 1 - NONPARAMETRIC DEFINITION & IDENTIFICATION OF THE CAUSAL EFFECTS

Traditionally, mediation analysis has been formulated, understood, and implemented within a framework of linear regression models. Unfortunately, this line of thinking cannot offer general definitions of the causal effects beyond a few specific models and its conclusions cannot be generalised to nonlinear models for discrete measures. In response, researchers have proposed the counterfactual framework to include the definition, identification, and estimation of causal mediation effects without any reference to one specific statistical model (VanderWeele and Vansteelandt, 2009; Pearl, 2010; Imai et al., 2010a; Pearl, 2012).

3.1 The counterfactual framework

Before we introduce a nonparametric definition for the causal effects, let us explain the concept of 'counterfactual outcomes' in settings where all variables are measured at the lower-level. A 'counterfactual' or 'potential' outcome $Y_{ij}(x)$ represents the outcome that we would, possibly contrary to fact, have observed for measurement j within cluster i, had the exposure X_{ij} been manipulated to a value x (Rubin, 1978). When considering a dichotomous exposure (with value 0 for baseline/no exposure, and 1 otherwise), we can define two possible potential outcomes for each measurement within a cluster: $Y_{ij}(0)$ and $Y_{ij}(1)$. Keeping this in mind, the measure- and cluster-specific total effect of X on Y is defined as the difference between both counterfactuals: $Y_{ij}(1) - Y_{ij}(0)$. Unfortunately, since only one of these counterfactuals is observed for each measurement, this effect cannot be estimated. The population average of the total causal effect $E[Y_{ij}(1) - Y_{ij}(0)]$, on the other hand, can be identified under specific assumptions (cfr. next section).

Similarly, counterfactuals for the mediator, $M_{ij}(0)$ and $M_{ij}(1)$, and nested counterfactuals for the outcome, $Y_{ij}(x, M_{ij}(x^*))$, can be devised (Robins and Greenland, 1992; Pearl, 2001). The latter counterfactual represents the value for the outcome Y_{ij}, when X_{ij} is set to x and M_{ij} is fixed at the value it would obtain when $X_{ij} = x^*$. Nested counterfactuals allow us to rephrase the average total effect of X on Y to include a mediator: $E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))] = E[Y_{ij}(1) - Y_{ij}(0)]$. This enables us to partition the total causal effect (TCE) into a total natural indirect (TNIE) and a pure natural direct effect (PNDE) (Hafeman and Schwartz, 2009; VanderWeele, 2013)[2]:

$$
\begin{aligned}
(1) \qquad TCE =& E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))] \\
=& E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0)) \\
&+ Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))] \\
=& TNIE + PNDE
\end{aligned}
$$

3.2 Causal and modelling assumptions to identify the causal mediation effects

In order to identify the above-defined non-parametrical effects in randomised lower-level mediation settings, we need to postulate the following set of assumptions (VanderWeele, 2010b; Josephy et al., 2015):

- *(i)* There are no unmeasured upper- or lower-level confounders of mediator and outcome.
- *(ii)* There are no confounders of mediator and outcome, caused by exposure.
- *(iii)* There is no carry-over effect when lower-level measures represent time points.

For lower-level mediation in clusters of size two, assumptions *(i)-(iii)* can be summarised by the (lack of) arrows within the diagram in Figure 1 (Robins and Richardson, 2010).

[1] We use data from 56 out of the original 68 couples, running a complete case analysis, as missingness proves problematic for the joint approach implemented through MPLUS®-software.

[2] Note that we define non-parametrical expressions for the direct and indirect effect on a linear scale, in contrast to e.g., VanderWeele (2013); Loeys et al. (2013); Bind et al. (2016).

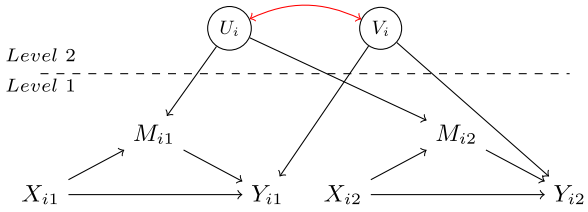

Figure 1. This causal diagram represents assumptions (i)-(iii)*, which are needed to identify the causal effects in a randomised lower-level mediation setting with clusters of size two. X_{i1}, M_{i1} and Y_{i1} represent the respective values of the exposure, mediator, and outcome for the first measure within cluster i, while X_{i2}, M_{i2} and Y_{i2} reflect these variables for the second measurement. Absence of a unidirectional arrow indicates the absence of a direct causal effect, while a bidirectional arrow captures an unmeasured common cause.*

Note however that including the red arrow in Figure 1, allows for a correlation between the unmeasured cluster-specific common causes of the outcome (V) and those of the mediator (U). Because both unmeasured confounders are correlated, we are able to define one confounder in terms of the other, without a loss of generality. For example, we can define an unspecified function h such that $V = h(U)$, which enables us to express the upper-level confounder V as a function of U, as to explicate their correlation. Consequently, this arrow directly violates assumption *(i)*: there *are* unmeasured *upper-level* confounders of the mediator-outcome relation. Josephy et al. (2015) showed that in linear lower-level mediation settings, this assumption is not necessary for the identification of the causal effects. In this manuscript, we wish to additionally demonstrate the redundancy of this assumption for binary lower-level mediation.

In addition to these three causal assumptions, we will consider the following modelling assumptions throughout the paper:

(iv) Unmeasured upper-level confounders of mediator and outcome exert an additive effect on both variables[3].

(v) There is no unmeasured heterogeneity among clusters in the effect of exposure on mediator, nor in the effect of exposure and mediator on the outcome.

Unlike the previous three assumptions, assumptions *(iv)* and *(v)* cannot be represented on a causal diagram; hence, they are not depicted in Figure 1.

4. STEP 2 - PARAMETRIC IDENTIFICATION OF THE CAUSAL EFFECTS

Now that we possess non-parametric definitions of the causal effects, we can pursue their identification based on parametrical models for our binary mediator and outcome. Let us consider the following multilevel models, with i the cluster, and j a within-cluster observation:

$$
\begin{aligned}
E[M_{ij}|X_{ij}, U_i] &= g_M^{-1}\big(\delta_M + \alpha X_{ij} + U_i\big) \\
E[Y_{ij}|X_{ij}, M_{ij}, U_i] &= g_Y^{-1}\big(\delta_Y + \zeta' X_{ij} + \beta M_{ij} \\
&\qquad + \phi X_{ij} M_{ij} + h(U_i)\big)
\end{aligned} \tag{2}
$$

where g_M^{-1} and g_Y^{-1} represent known inverse link functions for mediator and outcome, respectively. In these equations, δ_M and δ_Y represent the intercepts, while α, β, ζ', and ϕ represent the effects of exposure on mediator, mediator on outcome, exposure on outcome, and the interaction between exposure and mediator on the outcome, respectively. Note that we additionally assume that these effects are homogeneous across clusters, in accordance with assumption *(v)*. Since the unmeasured upper-level confounders of the mediator (U_i) and the outcome ($h(U_i)$) are allowed to correlate, this induces unmeasured cluster-specific confounding of the M-Y relationship (see red arrow in Figure 1).

Under this data-generating mechanism, all assumptions introduced in section 3.2 are met (except for the upper-level confounders of assumption *(i)*, of which we aim to prove its redundancy under a lenient set of modelling assumptions). This enables us to operate Pearl's mediation formula (Pearl, 2001, 2010) to derive the total, pure natural direct, and total natural indirect effect for each measurement j within cluster i.

For example, when $g_M = g_Y = probit$ (and hence with Φ representing the standard normal cumulative distribution), we find a "ij-th"-specific *total natural indirect effect*:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i)] \\
&\quad = \big(\Phi(\delta_M + U_i) - \Phi(\delta_M + \alpha + U_i)\big)\big(\Phi(\delta_Y + \zeta' + h(U_i)) \\
&\quad - \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\big)
\end{aligned} \tag{3}
$$

The parametrical derivations and expressions for the causal effects can be found in the appendix, for both $g_M = g_Y = logit$ and $g_M = g_Y = probit$.

5. STEP 3 - ESTIMATION MODELS FOR THE MEDIATOR AND OUTCOME

Now that we know how to identify the causal mediation effects, a next logical step aims to predict the counterfactuals. Before we can achieve this, however, we first need to estimate the regression parameters in equation (2) with the aid of appropriate estimation models. To this end, the following sections summarise three potential approaches[4].

[3]This assumption is made on the scale of the parametrical models for mediator and outcome

[4]Note that the next equations represent estimation models, in contrast to causal model (2) from the previous section.

5.1 Separate modelling of a binary mediator and outcome

One approach fits the mediator and outcome measures through separate multilevel models:

$$
\begin{aligned}
(4)\quad & E(M_{ij}|X_{ij},u_i) = g_M^{-1}\big(d_M + aX_{ij} + u_i\big)\\
& E(Y_{ij}|X_{ij},M_{ij},v_i)\\
& \quad = g_Y^{-1}\big(d_Y + c'X_{ij} + bM_{ij} + fX_{ij}M_{ij} + v_i\big)\\
& \quad \text{with } u_i \amalg X_{ij} \text{ and } v_i \amalg X_{ij}, M_{ij}, X_{ij}M_{ij}
\end{aligned}
$$

Here, u_i and v_i represent the random intercepts for M and Y, respectively. These upper-level residuals are assumed to be normally distributed with mean zero and variance σ^2_M for u_i and σ^2_Y for v_i. Note that this uncentred (UN) separate modelling approach assumes that the upper-level residuals are independent of the predictors. If, however, there is upper-level confounding of the M-Y relation, both random intercepts will be correlated and (because u_i predicts M_{ij}) M_{ij} and v_i will be correlated as well. This is in direct violation of the assumption that $v_i \amalg X_{ij}, M_{ij}$ and, as a result, the model for the outcome will estimate the regression coefficients of equation (2) with bias.

In linear multilevel settings (i.e., when g_M and g_Y both represent the identity-link), this issue can be tackled by separating within- from between-cluster effects (Louis, 1988; Neuhaus and Kalbfleisch, 1998; Begg and Parides, 2003; Zhang et al., 2009; Kenward and Roger, 2010; Preacher et al., 2010; Pituch and Stapleton, 2012). Such centering within-clusters (CWC) can be achieved by regressing a continuous dependent variable on the cluster-mean centred values of the predictors: $(X_{ij} - \overline{X}_i)$ and $(M_{ij} - \overline{M}_i)$. In these expressions, $\overline{X}_i$ and $\overline{M}_i$ denote the averages of the exposure- and mediator-scores within cluster i (MacKinnon, 2008). Subtracting these means from the raw scores will remove any cluster-specific effects that may influence the predictors and hence, any possible impact of unmeasured upper-level confounders. Consequently, the parameter coefficients of equation (2) may be estimated without bias in the presence of upper-level endogeneity, as the upper-level residuals will not correlate with the within-cluster deviations.

A similar approach is possible for a binary outcome:

$$
\begin{aligned}
(5)\quad & E(M_{ij}|X_{ij},u_i) = g_M^{-1}\big(d_M + a(X_{ij} - \overline{X}_i) + u_i\big)\\
& E(Y_{ij}|X_{ij},M_{ij},v_i)\\
& \quad = g_Y^{-1}\big(d_Y + c'(X_{ij} - \overline{X}_i) + b(M_{ij} - \overline{M}_i)\\
& \qquad + f(X_{ij}M_{ij} - \overline{XM}_i) + v_i\big)\\
& \quad \text{with } u_i \amalg (X_{ij} - \overline{X}_i) \text{ and } v_i \amalg (X_{ij} - \overline{X}_i),\\
& \quad (M_{ij} - \overline{M}_i), (X_{ij}M_{ij} - \overline{XM}_i)
\end{aligned}
$$

Again, both upper-level residuals are assumed to be independently and normally distributed with mean zero and fixed variance. Unfortunately, when a binary outcome is modelled on the *logit*- or *probit*-scale, CWC no longer yields proper parameter estimates for the within- and between-cluster effects (although in practice this bias may often be small (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010)).

5.2 Joint modelling of a binary mediator and outcome

A second approach jointly models the mediator and outcome. This can be achieved by either relying on multivariate techniques, or by tricking univariate software into modelling the mediator and outcome in a multivariate way (Bauer et al., 2006). The set of equations are identical to model (4) except that now, u_i and v_i are allowed to covary: $(u_i, v_i) \sim N(\mathbf{0}, \mathbf{\Sigma})$.

Here, the upper-level residuals are assumed to be multivariate normally distributed, with zero mean and covariance matrix $\mathbf{\Sigma}$. This matrix is defined by the variances of u_i and v_i on its diagonal (τ^2_M and τ^2_Y, respectively), and by the covariance between both upper-level residuals (τ_{MY}) elsewhere. Since this set of models allow both random intercepts to covary, unmeasured upper-level M-Y confounding may be accounted for through the modelling of this correlation. As this estimation model equals the true data-generating model (2) from section 4, we expect unbiased estimators for the regression coefficients in the outcome model.

6. STEP 4 - ESTIMATION OF THE CAUSAL EFFECTS THROUGH MONTE CARLO POTENTIAL OUTCOME GENERATION

Once an expression is derived for the direct or indirect effect of interest (such as expression (3)), we could simply plug in the parameter estimates for the fixed and random effects of the above described estimation models. However, this approach requires a new derivation for the expressions of the causal effects each time a different outcome or mediator model is considered. This can be remedied through a Monte-Carlo approach, which we explain in the following paragraph.

Recall that for a randomised binary exposure X, we observe $Y_{ij}(X_{ij}, M_{ij}(X_{ij}))$ for each within-cluster measure. However, in order to estimate the population averaged indirect effect, we also require the counterfactual outcome $Y_{ij}(X_{ij}, M_{ij}(1 - X_{ij}))$ for every measurement. In an algorithm proposed by Imai et al. (2010a), we can obtain a Monte Carlo draw from the potential outcome $Y_{ij}(x, M_{ij}(x^*))$ by using model predictions:

A Fit models for the observed mediator and outcome variables, such as described in section 5.
B Simulate estimated model parameters from their sampling distributions (e.g., 1000 draws).
C Repeat the following three processes within a single draw from step B: (i) predict both potential values of

the mediator ($M_{ij}(0)$ and $M_{ij}(1)$) for each measure within each cluster, (ii) predict the potential outcomes for each within-cluster measurement given the predicted values of the mediator ($Y_{ij}(0, M_{ij}(0))$, $Y_{ij}(1, M_{ij}(0))$, $Y_{ij}(0, M_{ij}(1))$, and $Y_{ij}(1, M_{ij}(1))$), (iii) compute the causal mediation effects, averaged over clusters and measurements within clusters.

D Compute summary statistics such as point estimates and confidence intervals, over all simulated draws.

7. ESTIMATION TECHNIQUES AND SOFTWARE IMPLEMENTATIONS

Estimation models aside, there exist a lot of different estimation techniques and software implementations that allow us to fit the above-mentioned statistical models and generate potential outcomes. Let us briefly go over several options.

7.1 Step 3 - Estimation of the regression parameters

Josephy et al. (2016) concluded that generalised linear mixed models (GLMMs) that rely on Maximum Likelihood (ML) estimation through Adaptive Gaussian Quadrature (AGQ) provided the most reliable estimates when analysing binary *probit*-regression models within small clusters. For dyadic cluster sizes, AGQ operated on par with Diagonally Weighted Least Squares (DWLS) estimation, but took the upper hand as the cluster size increased. With a preference for AGQ or DWLS as estimation techniques, our next obstacle aims to identify appropriate software implementations for our modelling approaches.

Let us take a look at possible implementations that allow us to jointly model the mediator and outcome. Bauer et al. (2006) first introduced a joint approach for linear mixed models (LMMs) in `SAS`® by fitting a multivariate model relying on the univariate multilevel Proc Mixed procedure. Extending this line of thinking to GLMMs, would require the specification of random effects in combination with a residual covariance structure that differentiates mediator from outcome. Unfortunately, Proc Glimmix cannot integrate marginal covariances within AGQ, while Proc NLmixed is unable to model residual covariances in the first place.

Structural Equation Models (SEM), where a categorical outcome is considered a crude approximation of an underlying continuous variable, offer a second possibility. As SEM naturally considers data in a multivariate way, it allows the joint modelling of mediator and outcome. Within (D)WLS-estimation, a binary measure that simultaneously acts as both a dependent and independent variable (i.e., an endogenous variable), is treated as its underlying continuous measure during the entire estimation process. As such, (D)WLS encounters problems when estimating the parameter coefficient of a binary mediator within the outcome model. Fortunately, ML-estimation treats the mediator as its underlying measure when it serves as a dependent variable, while considering its observed values when the mediator serves as a predictor. In contrast to Rosseel (2012)'s *lavaan*-package within `R`, where only (D)WLS-estimation is currently able to deal with endogenous categorical variables, `MPLUS`®-software can model such mediators values through ML-estimation with AGQ (Muthén and Muthén, 2010).

With these considerations and limitations in mind, we will consider ML-estimation through AGQ for the joint and separate modelling methods in the third step. The uncentred and centred separate modelling approaches will be fitted with the aid of the *lme4*-package (version 1.1-17) within `R` version 3.5.0 (Bates et al., 2015), while the joint approach will take place within the `MPLUS`-software (version 7.4).

7.2 Step 4 - Estimation of the causal mediation effects

Conveniently, Tingley et al. (2014) developed the `R`-package *mediation* in which separate models for M and Y can be inserted to generate estimates for the causal mediation effects with the aid of the algorithm described in section 6. We do, however, have a few concerns regarding the implementation as described in Imai et al. (2010a).

First, as this package can only model the mediator and outcome separately, it cannot quantify any unmeasured upper-level confounding of M and Y through a covariance term between both random intercepts. As a consequence, the random effects will not be able to appropriately deal with upper-level endogeneity of mediator and outcome.

Two, the authors do not provide any recommendations concerning which estimation techniques ought to be used in which settings, and hence do not explicitly recommend the use of AGQ when the fitted models for the mediator and/or outcome constitute GLMMs. As such, uninformed researchers might not be aware that they are relying upon the Laplace approximation by default and consequently, shoulder the approach's shortcomings when dealing with non-normal data within small clusters (Tuerlinckx et al., 2006; Josephy et al., 2016).

Three, in the third step of their algorithm (see section 6), the authors rely on a conditional approach for generating the random effects: the upper-level residuals are assumed to follow a normal distribution, conditional on the estimated random effect within that cluster and the estimated conditional variance. It has been pointed out that this method may not lead to a realistic sampling distribution (Skrondal and Rabe-Hesketh, 2004). Rather, a marginal sampling process, in which all random effects are drawn from a normal distribution with a zero mean and a standard deviation based on the estimated variance component, may lead to better estimates of the causal mediation effects. In response, we compare the performance of two possible random effect generating mechanisms within the fourth estimation step: a conditional vs. a marginal procedure. As `MPLUS` is currently unable to provide conditional variances,

Table 1. A summary of the parameter values for the data-generating mechanism, according to one of two possible link functions

	Link-function	
Parameter	*Probit*-link	*Logit*-link
δ_M	0.00	0.00
α	1.00	1.70
δ_Y	−0.70	−1.20
ζ'	0.50	0.85
β	0.80	1.35

we only compare both approaches within the two separate modelling techniques.

8. SIMULATION STUDY

In the following sections, we compare five different approaches for estimating the causal mediation effects: (1) an uncentered (UN) and (2) centred separate (CWC) approach, and (3) a joint modelling procedure (Joint) with marginal random effects ("-Marg"), alongside (4) an uncentered (UN) and (5) a centred separate (CWC) approach with conditional random effects ("-Marg"). For example, we will refer to the Joint approach with a marginal random effects generation as "Joint-Marg". The detailed code on software implementation can be found in the appendix.

To inspect the performances of these procedures, we compare them through simulations. For this, we generated binary mediator and outcome values within small clusters, according to random intercept *probit*- or *logit*-models (a summary of the parameter values can be found in Table 1). For simplicity, our data generating mechanism omits an interaction between exposure and mediator in the outcome model:

$$
\begin{aligned}
(6)\quad P(M_{ij}=1|X_{ij},u_i) &= P(M^*_{ij}>0|X_{ij},u_i)\\
&= P(\delta_M+\alpha X_{ij}+u_i+\epsilon^M_{ij}>0)\\
P(Y_{ij}=1|X_{ij},M_{ij},v_i)&\\
&= P(Y^*_{ij}>0|X_{ij},M_{ij},v_i)\\
&= P(\delta_Y+\zeta' X_{ij}+\beta M_{ij}+v_i+\epsilon^Y_{ij}>0)\\
&\text{with } (u_i,v_i)\sim N(\mathbf{0},\mathbf{\Sigma})
\end{aligned}
$$

Here, M^*_{ij} and Y^*_{ij} represent the underlying latent variables of the binary mediator and outcome, respectively, such that $M_{ij}=1$ if $M^*_{ij}>0$, and $Y_{ij}=1$ if $Y^*_{ij}>0$. In these equations, the lower-level residuals of the latent variables, ϵ^M_{ij} and ϵ^M_{ij}, are both i.i.d. drawn from a normal distribution with mean zero and a variance, σ^2, that changes according to the link function (for the *probit*-link, $\sigma^2=1$, for the *logit*-link $\sigma^2=\frac{\pi^2}{3}$). The random intercepts are sampled from a multivariate normal distribution with zero means and covariance matrix $\mathbf{\Sigma}$:

$$
\mathbf{\Sigma}=\begin{pmatrix}\tau_M & \rho\sqrt{\tau_M\tau_Y}\\ \rho\sqrt{\tau_M\tau_Y} & \tau_Y\end{pmatrix}
$$

For the different simulation settings, we vary several parameters: we consider different clusters sizes (2 vs. 5), a varying number of clusters (sample size $n=50$, 100, 300), two link functions for generating the mediator and outcome measures (*probit*- and *logit*-link)[5], the presence or absence of unmeasured upper-level confounding of mediator and outcome ($\rho=0$ vs. $\rho=0.50$), and three different intracluster correlations for the latent response variables. As the latent icc_l is defined as the proportion of between-cluster versus total variance in the latent responses (e.g., for the mediator, $icc_l=\frac{Var(u_i)}{Var(M^*_{ij})}=\frac{\tau_M}{\tau_M+\sigma^2}$), this value depends upon the variance of the lower-level residuals and hence, on the link function. As such, a latent icc_l of 0.10, 0.30, and 0.50 corresponds to a respective random intercept variance of 0.11, 0.43, and 1.00 for the *probit*-link, and of 0.36, 1.41, and 3.29 for the *logit*-link (with $icc^M_l=icc^Y_l$).

In total, 1000 simulations are generated for different combinations of cluster size, sample size, link-function, icc_l, and ρ. The five above-introduced methods are compared over these settings in terms of convergence, relative bias, mean squared error (MSE), and coverage. The relative bias is defined as the averaged difference between the estimated (e.g. $\hat{\beta}$) and true parameter values (e.g. β), divided by the latter (so that the relative bias equals $\frac{\beta-\hat{\beta}}{\beta}$); as such, a relative bias enclosing zero indicates an unbiased estimator. The MSE is estimated by summing the empirical variance and the squared bias of the estimates, simultaneously assessing bias and precision: the lower the MSE, the more accurate and precise the estimator. The coverage is defined as the proportion of the 95% Wald-confidence intervals that encompass their true parameter value; coverage rates nearing 95% represent nominal coverages of the intervals. Lastly, in order to conclude model convergence, a model fit must yield both estimates and standard errors. To ensure a fair comparison between methods, we only present results for simulation runs in which all five methods converge.

9. RESULTS

Below, we discuss the results of the simulation study for the *probit*-link in detail, and for clusters of size two and five. In addition, we report the results comparing the *logit*- and *probit*-link for clusters of size two.

9.1 Convergence

Here, we compare the results of three rather than five approaches, since both ways of generating the random intercepts overlap up until step 4 of our estimation process; hence, their convergence is identical. Generally, convergence improves as the sample size n and the number of measurements within a cluster increase (see left part of Figure 2). Note that for 300 clusters most approaches reach 100% convergence, except for the joint approach when the icc_l is low.

[5]Note that the coefficients of the *logit*-link are about 1.7 times larger compared to those defined for the *probit*-link (see Table 1).

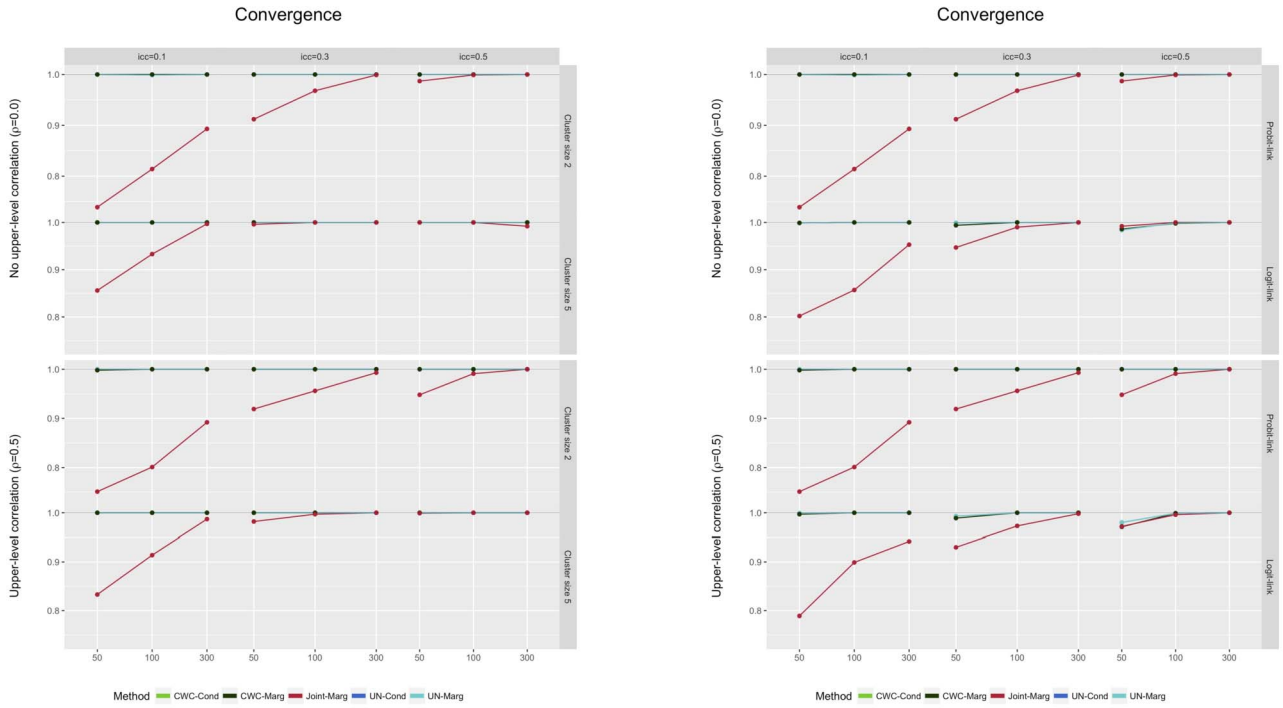

Figure 2. Model convergence of the five approaches comparing cluster sizes for the probit-link (left), and comparing link-functions for cluster size two (right), with different icc_l's (0.1, 0.3 and 0.5) and sample sizes (50, 100 and 300).

In contrast to changes in sample and cluster size, convergence is rather unaffected by the presence of unmeasured upper-level confounding of the mediator-outcome relation. Moreover, convergence is unaffected by the intracluster correlation for UN and CWC, whereas it seems to improve for the joint approach with increasing icc_l. Lastly, convergence fares slightly better for all approaches when the *logit*-link, rather than the *probit*-link is used (see right part of Figure 2). Overall, the joint approach shows the most difficulty in reaching convergence.

9.2 Relative bias

First of all, for the direct and indirect effect estimators we typically observe that the relative bias decreases as the number of clusters and the number of measurements within each cluster increases (see Figure 3). Only when there is upper-level endogeneity of mediator and outcome, does the relative bias of the indirect effect increase instead of decrease with larger samples, for both uncentred approaches. Second, both causal mediation effects are not influenced by an increase in the icc_l for the joint approach (with 'CWC-Cond' a close second), while it does impact others, especially when $\rho \neq 0$: in this case, their relative bias increases with rising *icc*. Third, when comparing link functions, we see no obvious changes in the performance of the joint approach, nor for both conditional approaches to generating the random effects (see Figure 4). Both marginal approaches, however, exhibit a strong increase in relative bias when relying on the *logit*-, compared to the *probit*-link. Overall, we observe that the joint approach provides the least biased estimates.

9.3 MSE

The MSE declines with increasing sample size and number of within-cluster measures, as well as with a rising icc_l (see Figure 5). Furthermore, we do not observe any differences in MSE when comparing settings with and without unmeasured upper-level confounding of the M-Y relation, nor when comparing link functions. The only deviation from this consists of a slightly increased MSE for the direct effect estimator of both CWC approaches when comparing *logit*- to *probit* regression (see Figure 6). All settings considered, the MSE is generally lowest for the joint modelling approach and CWC.

9.4 Coverage

For both causal mediation effects, the coverage of their 95% confidence intervals is typically better when the cluster size equals two rather than five, and when the intracluster correlation is low (see Figure 7). This observation holds for all methods, although the joint approach and 'CWC-Cond' are least influenced by such changes. Additionally, when the icc_l is high, we often observe a decrease in coverage as the upper-level sample size increases, for both UN-approaches and 'CWC-Marg'. Also, the presence of unmeasured upper-level confounding of M and Y does not seem to impact the joint approach or 'CWC-Cond' that much, whereas the

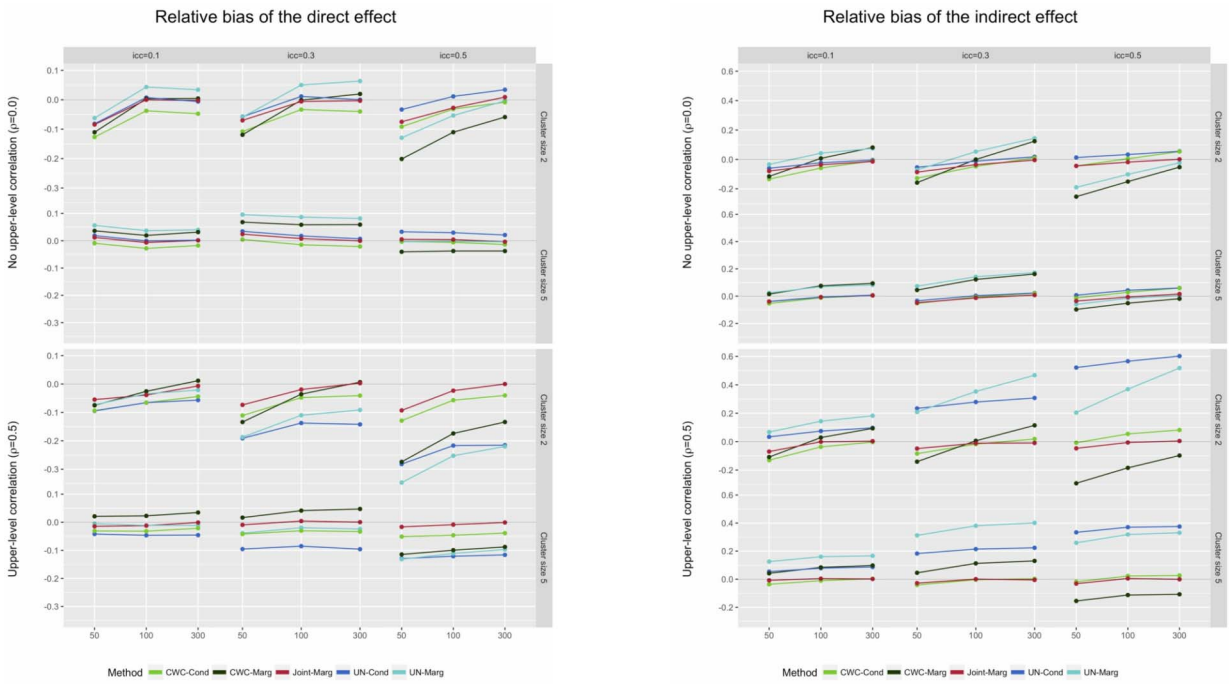

Figure 3. The relative bias of the direct (left) and indirect (right) for the five approaches modelled with a probit-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

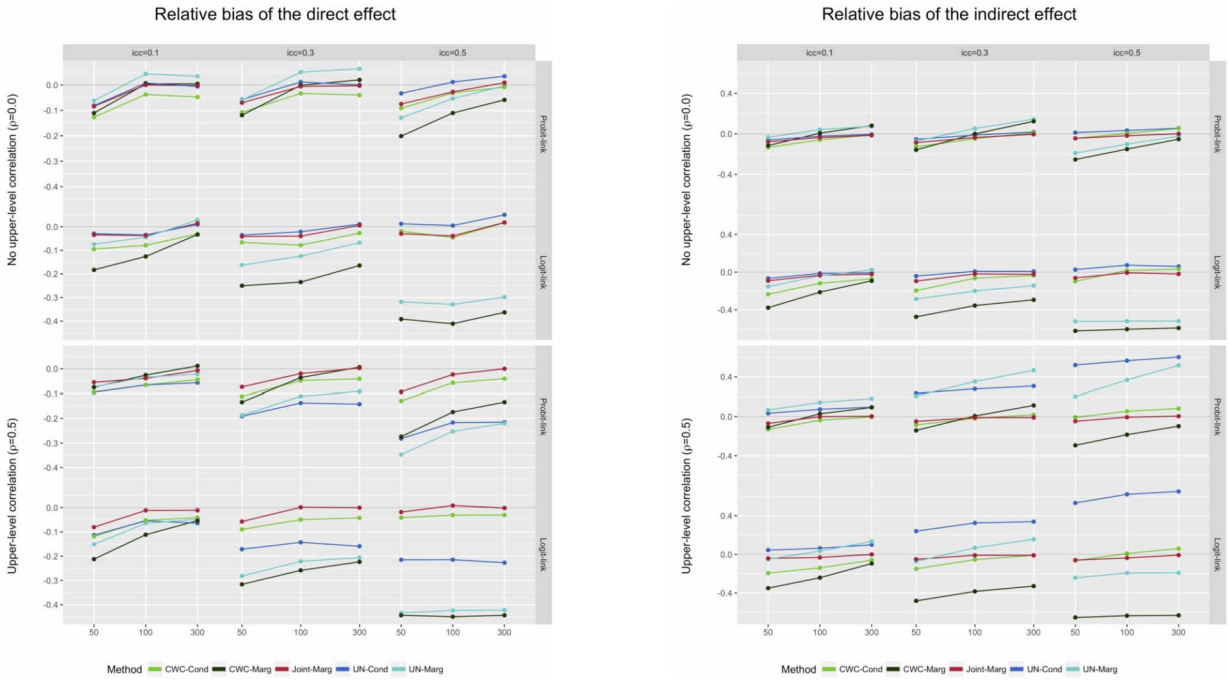

Figure 4. The relative bias of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (probit and logit), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

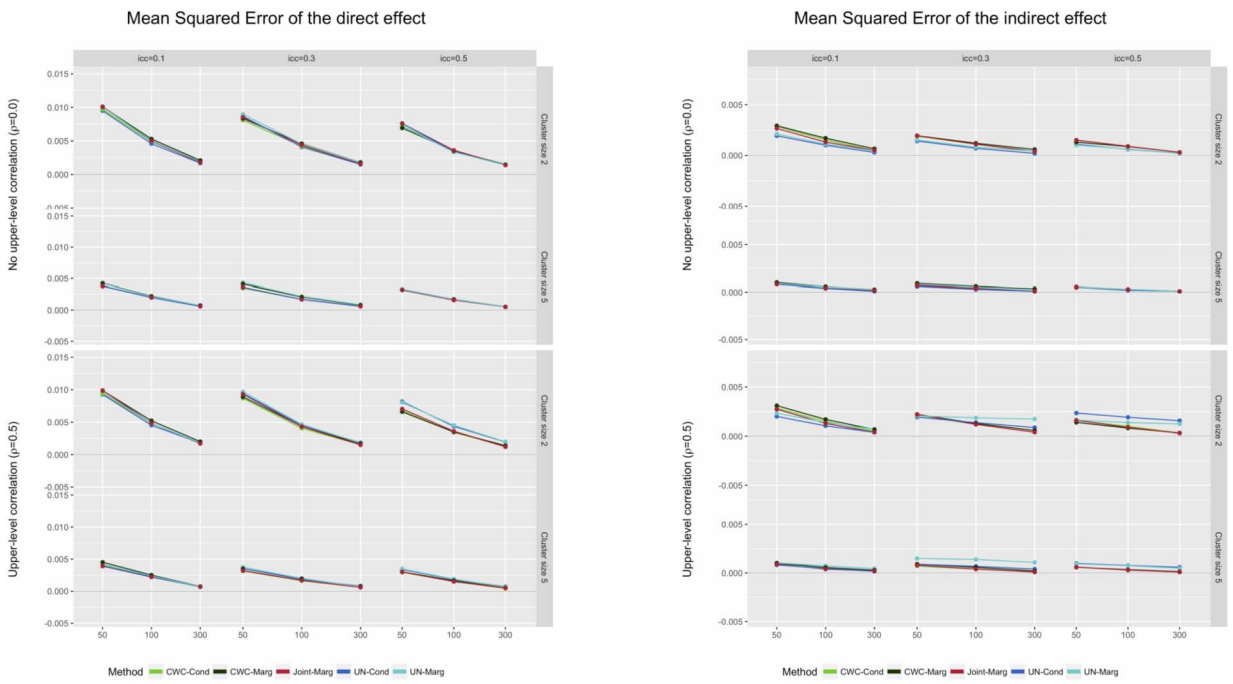

Figure 5. The MSE of the direct (left) and indirect (right) for the five approaches modelled with a probit-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

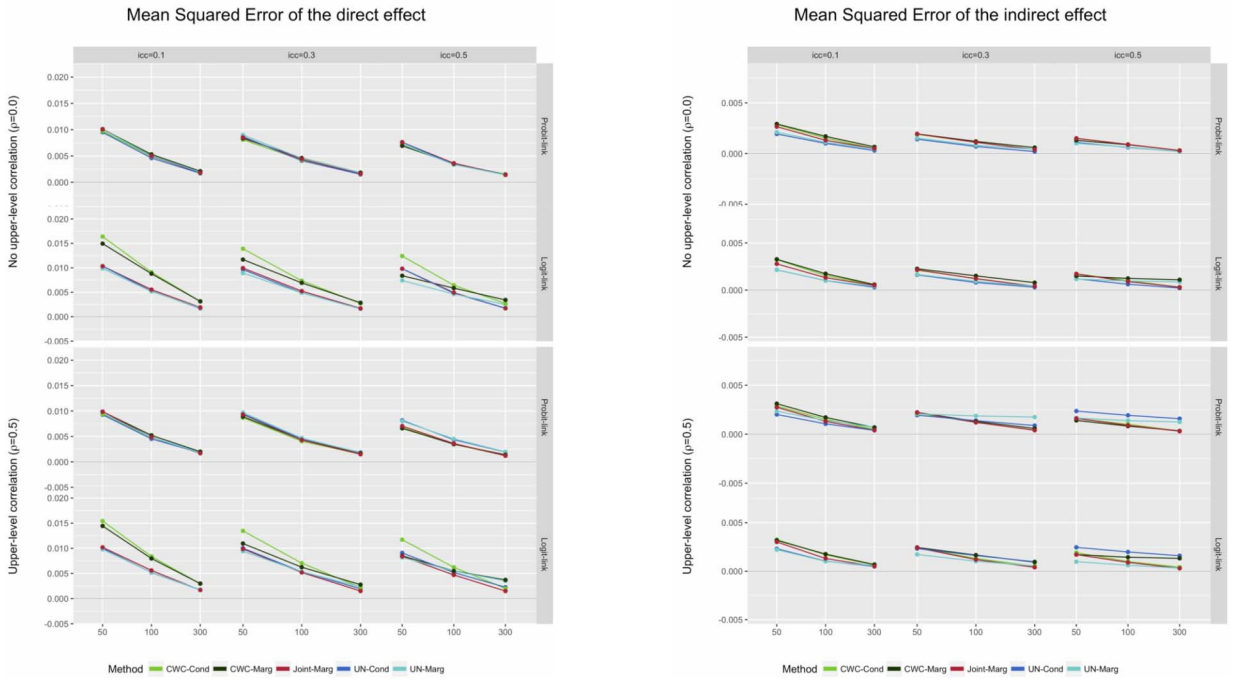

Figure 6. The MSE of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (probit and logit), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

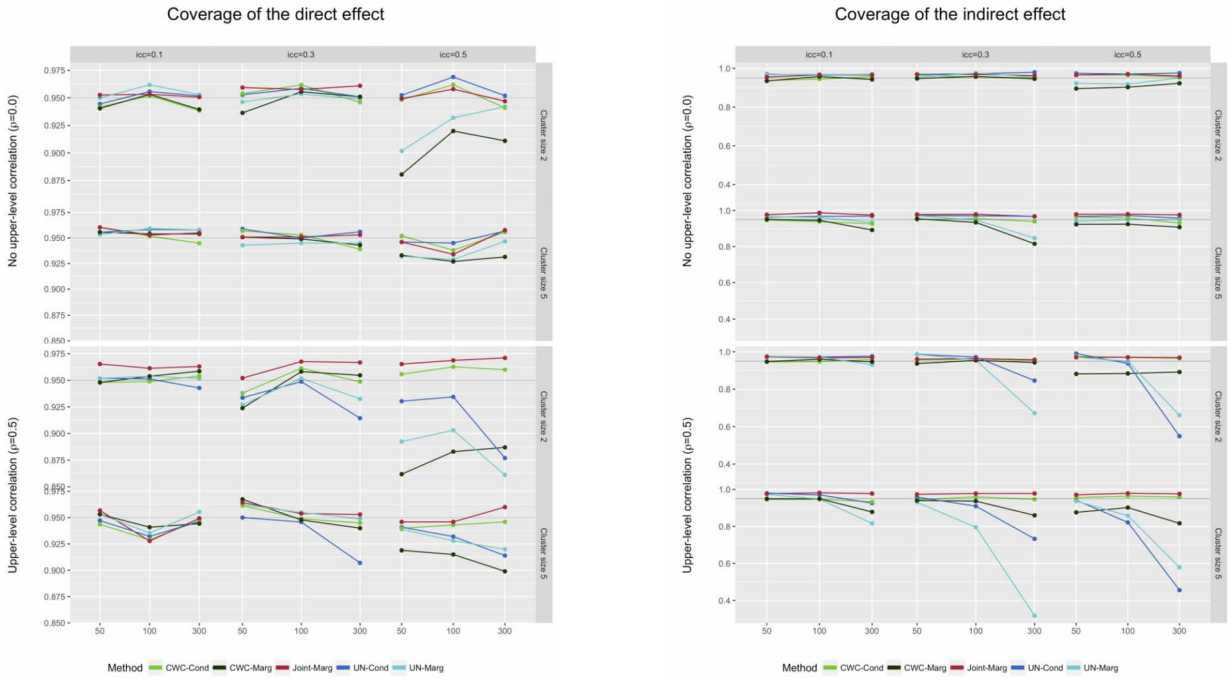

Figure 7. The coverage of the direct (left) and indirect (right) for the five approaches modelled with a probit-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

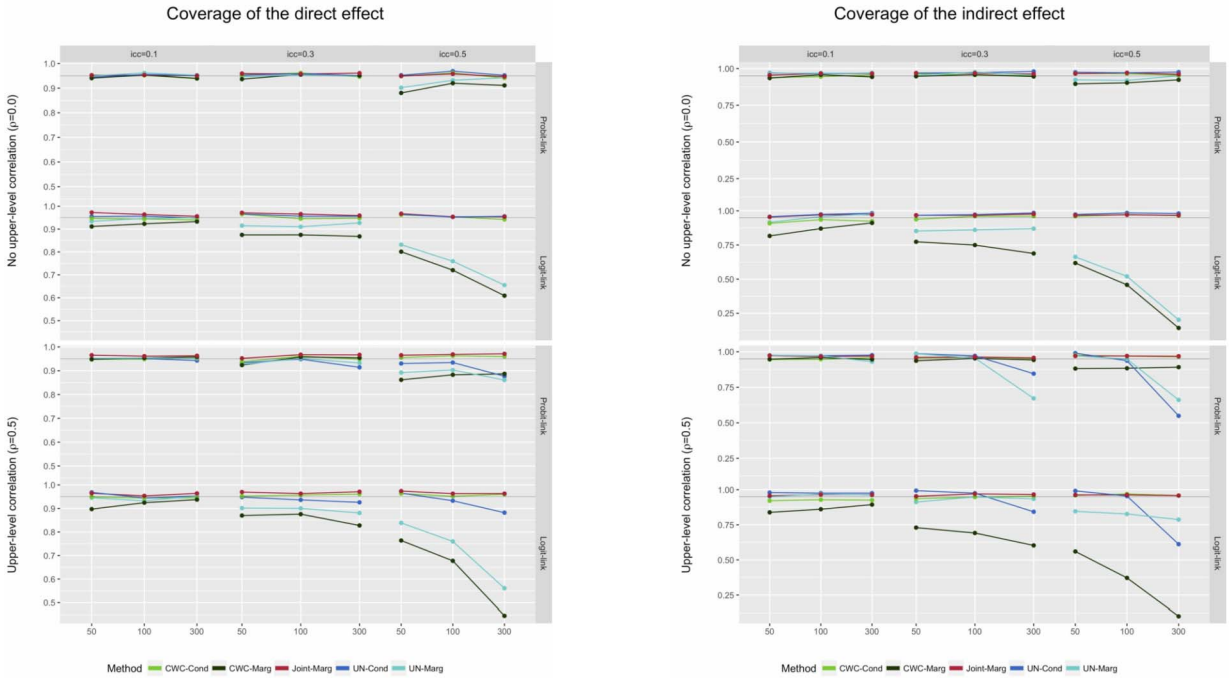

Figure 8. The coverage of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (probit and logit), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

Table 2. The estimates of the indirect, direct, and total effect of goal conflict on the observed amount of help, mediated by the partners' amount of autonomous helping motivation. The estimates (with empirical standard errors, se) and percentile-based 95% confidence intervals are provided for the five estimation methods

	Indirect Effect		Direct Effect		Total Effect	
	Estimate (se)	95%-CI	Estimate (se)	95%-CI	Estimate (se)	95%-CI
UN-Cond	-0.009 (0.015)	(-0.042; 0.016)	-0.390 (0.086)	(-0.561; -0.210)	-0.398 (0.087)	(-0.555; -0.212)
UN-Marg	-0.009 (0.015)	(-0.042; 0.015)	-0.409 (0.087)	(-0.589; -0.218)	-0.417 (0.091)	(-0.585; -0.230)
CWC-Cond	-0.002 (0.008)	(-0.020; 0.014)	-0.317 (0.112)	(-0.486; -0.131)	-0.318 (0.097)	(-0.492; -0.133)
CWC-Marg	-0.002 (0.008)	(-0.022; 0.012)	-0.324 (0.114)	(-0.518; -0.124)	-0.326 (0.107)	(-0.519; -0.122)
Joint-Marg	-0.005 (0.010)	(-0.029; 0.012)	-0.392 (0.093)	(-0.562; -0.212)	-0.397 (0.088)	(-0.564; -0.218)

other methods show a steep decrease in coverage, especially when samples sizes are large. Again, the link function does not seem to impact the coverage of the joint and both conditional approaches, whereas it tends to decrease for both marginal approaches when comparing the *logit*- to the *probit*-link (see Figure 8). Generally, the joint approach and 'CWC-Cond' provide the best coverage.

9.5 Summary

The convergence issues that the joint approach experiences are mostly limited to very small sample sizes (and even then never fall below 75%). When the joint model converges, however, there are numerous advantages to using the joint approach over others: it provides the least biased estimates, small standard errors (resulting in small MSE's), and offers coverages that are at least as good or better than the other approaches considered.

9.6 Analysis of the example

We illustrate these five approaches by applying them to our example data. Here, we assess whether of not the effect of goal conflict (the binary exposure) on a high vs. low amount of help (the binary outcome) is mediated by a high amount of autonomous helping motivation from the partner, as perceived by the patient (i.e., a high versus low amount of autonomous motivation, the binary mediator). Within the third step of our estimation procedure, we model the mediator and outcome according to the estimation models (4) and (5), where $g_M = g_Y = probit$ and with $f = 0$ (i.e., without an interaction).

This dataset demonstrates a small upper-level sample size ($n = 56$) with two measures in each cluster, a medium to small latent *icc*'s (0.32 for M and 0.16 for Y), and a high (nonsignificant) amount of unmeasured upper-level M-Y confounding (estimated at 0.78, with $p = 0.45$). In such settings, our simulation study suggests that (when the models are correctly specified) the joint approach is most likely to yield unbiased estimates for both mediation effects (as can be seen in Figure 3).

When goal conflict is present, we observe a total effect that indicates a decrease in the observed amount of help from high to low of about 32% to 42%. Additionally, we observe a significant direct, but no indirect effect for all five estimation procedures (see Table 2). For both effects, we see that the uncentred estimates are larger than those of both CWC-approaches. For the intervening effect, we can additionally observe that the estimates of the Joint approach lie somewhere in between the centred and uncentred approaches, while the estimates for the joint approach of the direct effect are very close to the estimates of both uncentred approaches. Overall, we can conclude that the partner's amount of autonomous helping motivation (a perceived by the patient) does not mediate the effect of goal conflict on the observed amount of help.

It is worth noting that although section 3.2 shows that assumption *(i)* can be relaxed to exclude upper-level confounders for some estimation models, the other assumptions still need to hold in order for our inferences to be valid. Unfortunately, there is no way of checking the plausibility of assumptions *(i)*), *(ii)*, and *(iv)*. Concerning assumption *(iii)*, we cannot exclude the possibility of a carry-over effect, as there was no notable wash-out period in this study. There is also no way of checking assumption *(v)*, as random slopes are unidentifiable in designs with a mere two measurements within each cluster. These restrictions should be kept in mind when interpreting results.

10. DISCUSSION

In this paper, we provided an overview of several possible estimation techniques that allow us to evaluate lower-level mediation in binary settings (with a specific focus on small cluster sizes). Additionally, we presented an extensive simulation study in which we assessed the impact of several data features on the convergence, relative bias, mean squared error, and coverage of the estimates of the various methods. We found that jointly modelling the mediator and the outcome provided the best performance measures (combined with a marginal approach to simulating the random effects), especially in the presence of unmeasured upper-level confounding of mediator and outcome. A separate modelling approach that centres the lower-level variables within-clusters and draws the random effects conditionally, comes in as a close second, confirming the reports of Goetgeluk and Vansteelandt (2008); Brumback et al. (2010). These authors stated that although

CWC no longer yields proper parameter estimates when the outcome is binary, the resulting bias is often small. Unsurprisingly, not centering the lower-level predictors provided very biased estimates in the presence of upper-level mediator-outcome endogeneity (irrespective of the assumed random effects distribution).

These conclusions in mind, we must acknowledge several limitations to this manuscript. For one, we restricted our simulations to settings where the intracluster correlations for mediator and outcome are identical, as allowing them to vary independently of each other would have incremented the computational demands by sixfold. However, varying the icc the particular setting where the *probit*-link was used in the presence of upper-level M-Y confounding and with clusters of size two, supported the conclusions summarised here (results not shown).

Two, there exists a non-parametrical implementation of the algorithm that we described in step 4 of the estimation process. This alternative assesses mediation based on bootstrapping mediator and outcome values, rather than on simulated draws from the estimated parameter distributions. We did not incorporate this procedure within our study (in accordance to Imai et al. (2010a)'s *mediation*-package), as the bootstrapping process takes up an enormous amount of time in multilevel samples. However, our (and Imai et al. (2010a)'s) sole reliance on a parametrical approach might provide suboptimal estimates for the causal mediation effects, especially when estimation procedures are used that may produce biased estimates (e.g., the uncentred approaches in the presence of upper-level endogeneity of M and Y).

Three, we only considered complete data in our simulations as missingness in either mediator or outcome will cause `Mplus` to produce error messages when the method of integration is specified as 'Gaussian' (i.e., as in Gaussian Adaptive Quadrature). In contrast, the two separate modelling approaches in `R` consider all available outcomes, even when there is missingness.

APPENDIX A. IDENTIFICATION OF CAUSAL EFFECTS

In the following sections, we derive the parametrical expressions for the causal effects when $g_M = g_Y = probit$, and when $g_M = g_Y = logit$.

A.1 Probit-regression models

Consider these *probit*-models for a binary M and Y (with i the cluster and j the measurement within a cluster):

$$
\begin{aligned}
&E[M_{ij}|X_{ij}, U_i] = \Phi\big(\delta_M + \alpha X_{ij} + U_i\big)\\
&E[Y_{ij}|X_{ij}, M_{ij}, V_i]\\
&\quad = \Phi\big(\delta_Y + \zeta' X_{ij} + \beta M_{ij} + \phi X_{ij} M_{ij} + h(U_i)\big)
\end{aligned}
$$

with Φ representing the standard normal cumulative distribution. Based on this data generating mechanism, the "ij-th"-specific *total natural indirect effect* can be identified, when the assumptions *(i)-(v)* from section 3.2 are satisfied:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i, t)])\\
&\quad = \sum_m \{E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]\\
&\qquad \times P(M_{ij} = m|X_{ij} = 1, U_i)\\
&\qquad - E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]\\
&\qquad \times P(M_{ij} = m|X_{ij} = 0, U_i)\}\\
&\quad = P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)\\
&\qquad \times (1 - P(M_{ij} = 1|X_{ij} = 1, U_i))\\
&\qquad - P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)\\
&\qquad \times (1 - P(M_{ij} = 1|X_{ij} = 0, U_i))\\
&\qquad + P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)\\
&\qquad \times (P(M_{ij} = 1|X_{ij} = 1, U_i)\\
&\qquad - P(M_{ij} = 1|X_{ij} = 0, U_i))\\
&\quad = P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)\\
&\qquad \times (P(M_{ij} = 1|X_{ij} = 0, U_i)\\
&\qquad - P(M_{ij} = 1|X_{ij} = 1, U_i))\\
&\qquad + P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)\\
&\qquad (P(M_{ij} = 1|X_{ij} = 1, U_i) - P(M_{ij} = 1|X_{ij} = 0, U_i))\\
&\quad = (P(M_{ij} = 1|X_{ij} = 0, U_i) - P(M_{ij} = 1|X_{ij} = 1, U_i))\\
&\qquad (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)\\
&\qquad - P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i))\\
&\quad = \big(\Phi(\delta_M + U_i) - \Phi(\delta_M + \alpha + U_i)\big)\big(\Phi(\delta_Y + \zeta' + h(U_i))\\
&\qquad - \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\big)
\end{aligned}
$$

as is the "ij-th"-specific *pure natural direct effect*:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))|U_i, t)]\\
&\quad = \sum_m \{E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]\\
&\qquad \times P(M_{ij} = m|X_{ij} = 0, U_i)\\
&\qquad - E[Y_{ij}|X_{ij} = 0, M_{ij} = m, U_i]\\
&\qquad \times P(M_{ij} = m|X_{ij} = 0, U_i)\}\\
&\quad = (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)\\
&\qquad - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 0, U_i))\\
&\qquad P(M_{ij} = 0|X_{ij} = 0, U_i)\\
&\qquad + (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)\\
&\qquad - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 1, U_i))\\
&\qquad \times P(M_{ij} = 1|X_{ij} = 0, U_i)\\
&\quad = \big(1 - \Phi(\delta_M + U_i)\big)\\
&\qquad \times \big(\Phi(\delta_Y + \zeta' + h(U_i)) - \Phi(\delta_Y + h(U_i))\big)\\
&\qquad + \Phi(\delta_M + U_i)\big(\Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\\
&\qquad - \Phi(\delta_Y + \beta + h(U_i))\big)
\end{aligned}
$$

and the "ij-th"-specific *total causal effect*:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&= \sum_m \{E[Y_{ij}|X_{ij}=1, M_{ij}=m, U_i] \\
&\quad \times P(M_{ij}=m|X_{ij}=1, U_i) \\
&\quad - E[Y_{ij}|X_{ij}=0, M_{ij}=m, U_i] \\
&\quad \times P(M_{ij}=m|X_{ij}=0, U_i)\} \\
&= P(Y_{ij}=1|X_{ij}=1, M_{ij}=0, U_i) \\
&\quad \times P(M_{ij}=0|X_{ij}=1, U_i) \\
&\quad - P(Y_{ij}=1|X_{ij}=0, M_{ij}=0, U_i) \\
&\quad \times (P(M_{ij}=0|X_{ij}=0, U_i)) \\
&\quad + P(Y_{ij}=1|X_{ij}=1, M_{ij}=1, U_i) \\
&\quad \times (P(M_{ij}=1|X_{ij}=1, U_i)) \\
&\quad - P(Y_{ij}=1|X_{ij}=0, M_{ij}=1, U_i) \\
&\quad \times (P(M_{ij}=1|X_{ij}=0, U_i)) \\
&= \Phi(\delta_Y + \zeta' + h(U_i))\big(1 - \Phi(\delta_M + \alpha + U_i)\big) \\
&\quad - \Phi(\delta_Y + h(U_i))\big(1 - \Phi(\delta_M + U_i)\big) \\
&\quad + \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\Phi(\delta_M + \alpha + U_i) \\
&\quad - \Phi(\delta_Y + \beta + h(U_i))\Phi(\delta_M + U_i)
\end{aligned}
$$

A.2 Logit-regression models

Consider these *logit*-models for a binary M and Y:

$$
\begin{aligned}
E[M_{ij}|X_{ij}, U_i] &= \frac{1}{1 + e^{-\delta_M - \alpha X_{ij} - U_i}} \\
E[Y_{ij}|X_{ij}, M_{ij}, U_i] &= \frac{1}{1 + e^{-\delta_Y - \zeta' X_{ij} - \beta M_{ij} - \phi X_{ij} M_{ij} - h(U_i)}}
\end{aligned}
$$

Based on this data generating mechanism, the "ij-th"-specific *total natural indirect effect* can be identified, when the assumptions *(i)-(v)* from section 3.2 are satisfied:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i, t)] \\
&\quad = \Big(\frac{1}{1 + e^{-\delta_M - U_i}} - \frac{1}{1 + e^{-\delta_M - \alpha - U_i}}\Big) \\
&\quad \times \Big(\frac{1}{1 + e^{-\delta_Y - \zeta' - h(U_i)}} - \frac{1}{1 + e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}}\Big)
\end{aligned}
$$

as is the "ij-th"-specific *pure natural direct effect*:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&= \frac{e^{-\delta_M - U_i}}{1 + e^{-\delta_M - U_i}}\Big(\frac{1}{1 + e^{-\delta_Y - \zeta' - h(U_i)}} - \frac{1}{1 + e^{-\delta_Y - h(U_i)}}\Big) \\
&\quad + \frac{1}{1 + e^{-\delta_M - U_i}}\Big(\frac{1}{1 + e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}} \\
&\quad - \frac{1}{1 + e^{-\delta_Y - \beta - h(U_i)}}\Big)
\end{aligned}
$$

and the "ij-th"-specific *total causal effect*:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&= \frac{1}{1 + e^{-\delta_Y - \zeta' - h(U_i)}} \frac{e^{-\delta_M - \alpha - U_i}}{1 + e^{-\delta_M - \alpha - U_i}} \\
&\quad - \frac{1}{1 + e^{-\delta_Y - h(U_i)}} \frac{e^{-\delta_M - U_i}}{1 + e^{-\delta_M - U_i}} \\
&\quad + \frac{1}{1 + e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}} \frac{1}{1 + e^{-\delta_M - \alpha - U_i}} \\
&\quad - \frac{1}{1 + e^{-\delta_Y - \beta - h(U_i)}} \frac{1}{1 + e^{-\delta_M - U_i}}
\end{aligned}
$$

APPENDIX B. SOFTWARE CODE

B.1 Data generating mechanism

For the generation of probit-data with cluster size two, sample size 'n', latent intracluster correlation 'icc', and random intercept-correlation between M and Y of 'rho':

```
#Generate 1000 data sets for the current n, icc, and rho:
for (i in 1:1000){
  #Population parameters:
  iM<-0; ia<-1; iY<--0.7; ic<-0.5; ib<-0.8
  #Random intercept covariance matrix (with tau<-icc/(1-icc)):
  sig<-matrix(c(tau,sqrt(tau)*sqrt(tau)*rho,
         sqrt(tau)*sqrt(tau)*rho,tau),byrow=T,nrow=2)
  #Random intercepts for M and Y within each cluster:
  ri<-mvrnorm(n,c(0,0),sig)
  #Generate data for binary X, M and Y:
  x0<-rbinom(n,1,0.5);  x1<-1-x0
  m0<-rbinom(n,1,pnorm(iM+ia*x0+ri[,1]))
  m1<-rbinom(n,1,pnorm(iM+ia*x1+ri[,1]))
  y0<-rbinom(n,1,pnorm(iY+ic*x0+ib*m0+ri[,2]))
  y1<-rbinom(n,1,pnorm(iY+ic*x1+ib*m1+ri[,2]))
  #Centring of X and M within-clusters:
  xmean<-colMeans(rbind(x0,x1));  xx0<-x0-xmean;  xx1<-x1-xmean
  mmean<-colMeans(rbind(m0,m1));  mm0<-m0-mmean;  mm1<-m1-mmean
  #Convert the variables to long format:
  x<-c(x0,x1); m<-c(m0,m1); xx<-c(xx0,xx1); mm<-c(mm0,mm1);
    y<-c(y0,y1)
  #Cluster identifier:
  ind<-rep(seq(1,n),2)
  #Create dataset:
  data<-as.data.frame(cbind(ind,x,m,xx,mm,y)) }
```

B.2 Estimation models

For uncentred separate modelling by use of `lme4` in `R` (Bates et al., 2015):

```
med.UN<-glmer(m~x+(1|ind),family=binomial(link="probit"),
        data=data,nAGQ=15)
out.UN<-glmer(y~x+m+(1|ind),family=binomial(link="probit"),
        data=data,nAGQ=15)
```

For CWC-separate modelling by use of `lme4` in `R` (Bates et al., 2015):

```
med.CWC<-glmer(m~xx+(1|ind),family=binomial(link="probit"),
        data=data,nAGQ=15)
out.CWC<-glmer(y~xx+mm+(1|ind),family=binomial(link="probit"),
        data=data,nAGQ=15)
```

For the joint modelling approach in `Mplus` (Muthén and Muthén, 2010):

```
DATA:  file = mplus.raw;  type = individual;
VARIABLE:  names = x0 x1 m0 m1 y0 y1;  usevariables = x0 x1
m0 m1 y0 y1;  missing = .;  categorical = m0 m1 y0 y1;
ANALYSIS:  type = general;  estimator = ML;  integration= GAUSS;
```

```
  adaptive = on;  link = probit;
MODEL:  i0 BY m0@1 m1@1;  i1 BY y0@1 y1@1;  i0 (Mvar);  i1 (Yvar);
  m0 ON x0 (a);  m1 ON x1 (a);
  y0 ON x0 (c);  y1 ON x1 (c);  y0 ON m0 (b);  y1 ON m1 (b);
  [m0$1] (iM);  [m1$1] (iM);  [y0$1] (iY);  [y1$1] (iY);
OUTPUT:  sampstat cinterval tech3;
```

B.3 Generation of the random effects

For marginally generated random effects, based on the uncentred separate modelling approach and the *probit*-link:

```
#Extract the estimates and estimated covariance matrix:
b_est<-c(fixef(med.UN),out.UN)
b_vcov[c(1:2),c(1:2)]<-as.matrix(vcov(med.UN))
b_vcov[c(3:5),c(3:5)]<-as.matrix(vcov(out.UN))
#Extract the estimated random intercept variances:
ri_varM<-med.UN@theta**2, ri_varY<-out.UN@theta**2
```

For conditionally generated random effects, based on the uncentred separate modelling approach and the *probit*-link:

```
#Extract the parameter estimates and estimated covariance matrix:
b_est<-c(fixef(med.UN),fixef(out.UN))
b_vcov[c(1:2),c(1:2)]<-as.matrix(vcov(med.UN))
b_vcov[c(3:5),c(3:5)]<-as.matrix(vcov(out.UN))
#Extract the estimated conditional means and random intercept vars:
ri_meanM<-ranef(med.UN)[[1]][,1];  ri_meanY<-ranef(out.UN)[[1]][,1]
ri_cond_varM<-cond.se(med.UN)[[1]][,1]
ri_cond_varY<-cond.se(out.UN)[[1]][,1]
#With the function to extract the conditional standard errors:
cond.se<-function(object){
  se.bygroup<-ranef(object,condVar=T)
  vars<-attr(se.bygroup[[1]],"postVar")
  se.by.clust[[1]]<-array(NA,c(n,1))
  for (j in 1:n){
    se.by.clust[[1]][j,]<-sqrt(diag(as.matrix(vars[,,j])))  }
  return(se.by.clust)}
```

Simulating draws from the sampling distributions, based on the probit-link and conditionally drawn random effects:

```
#Simulate draws from the sampling distribution:
b_sim<-mvrnorm(1000,b_est,b_vcov)
#Simulated draws:
for (t in 1:1000){
 riMc<-rep(rnorm(n,mean=ri_meanM,sd=ri_cond_varM),each=2)
 riYc<-rep(rnorm(n,mean=ri_meanY,sd=ri_cond_varY),each=2)
 m_0c<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMc))
 m_1c<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMc+b_sim[t,2]))
 y00c<-pnorm(b_sim[t,3]+riYc+b_sim[t,5]*m_0c)
 y11c<-pnorm(b_sim[t,3]+riYc+b_sim[t,4]+b_sim[t,5]*m_1c)
 y10c<-pnorm(b_sim[t,3]+riYc+b_sim[t,4]+b_sim[t,5]*m_0c)
 ie_c[t]<-mean(y11c-y10c); de_c[t]<-mean(y10c-y00c)
 te_c[t]<-mean(y11c-y00c)  }
#Conditional causal effects:
ie[i,1]<-mean(ie_c); de[i,1]<-mean(de_c); te[i,1]<-mean(te_c)
```

ACKNOWLEDGEMENTS

This research was supported by the Research Foundation Flanders (FWO) Grant G019317N.

Received 21 February 2019

REFERENCES

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2):135–167.

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11(2):142–163.

Begg, M. D. and Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, 22(16):2591–2602.

Bind, M. A. C., Vanderweele, T. J., Coull, B. A., and Schwartz, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics*, 17(1):122–134. MR3449855

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25. MR1397972

Brumback, B. A., Dailey, A. B., Brumback, L. C., Livingston, M. D., and He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters*, 80(21-22):1650–1654. MR2684013

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4):529–569.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780. MR2526627

Hafeman, D. M. and Schwartz, S. (2009). Opening the Black Box: A motivation for the assessment of mediation. *International Journal of Epidemiology*, 38(3):838–845.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71. MR2741814

Josephy, H., Loeys, T., and Rosseel, Y. (2016). A Review of R-packages for random-intercept probit regression in small clusters. *Frontiers in Applied Mathematics and Statistics*, 2(18):1–13.

Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., and Loeys, T. (2015). Within-subject mediation analysis in AB/BA crossover designs. *The International Journal of Biostatistics*, 11(1):1–22. MR3341510

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115–128.

Kenward, M. G. and Roger, J. H. (2010). The use of baseline covariates in crossover studies. *Biostatistics*, 11(1):1–17.

Kindt, S., Vansteenkiste, M., De Ruddere, L., Cano, A., and Goubert, L. (2018). "What should I do first?" The effect of manipulated goal conflict on affect, motivation and helping behavior in chronic pain. *Under review*.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, 48(6):871–894.

Louis, T. A. (1988). General methods for analyzing repeated measures. *Statistics in Medicine*, 7(1-2):29–45.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Taylor & Francis Group, LLC, New York.

McMahon, J. M., Tortu, S., Torres, L., Pouget, E. R., and Hamid, R. (2003). Recruitment of heterosexual couples in public health research: a study protocol. *BMC medical research methodology*, 3:24.

Montoya, A. K. and Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1):6–27.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, sixth edition.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Ortqvist, A. K., Lundholm, C., Carlström, E., Lichtenstein, P., Cnattingius, S., and Almqvist, C. (2009). Familial factors do not confound the association between birth weight and childhood asthma. *Pediatrics*, 124(4):e737–43.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article 7. MR2595111

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Pituch, K. a. and Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, 41(4):630–670. MR3190720

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(4):1–28.

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3):209–233.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and data analysis methods*. Sage, Thousand Oaks, CA, second edition.

Raykov, T. and Mels, G. (2007). Lower level mediation effect analysis in two-level studies: A note on a multilevel structural equation modeling approach. *Structural Equation Modeling*, 14(4):636–648. MR2413469

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):153–155.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A*, 158(1):73–89.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rovine, M. J. and Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1):55–88.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58. MR0472152

Senn, S. (2002). *Cross-over trials in clinical research*. John Wiley & Sons, Chichester.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York. MR2059021

Skrondal, A. and Rabe-Hesketh, S. (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 63(2):211–237. MR3234341

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage, Thousand Oaks, CA. MR3137621

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Tofighi, D. and Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behavioral Research*, 51(1):86–105.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *The British Journal of Mathematical and Statistical Psychology*, 59(2):225–255. MR2282214

VanderWeele, T. J. (2010a). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21(4):540–551.

VanderWeele, T. J. (2010b). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, 38(4):515–544. MR2758165

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468. MR2576399

Vuorre, M. and Bolger, N. (2017). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behavior Research Methods*, pages 1–19.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA. MR2768559

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models. *Organizational Research Methods*, 12(4):695–719.

Haeike Josephy
Department of Data Analysis
Ghent University
Henri Dunantlaan 2
9000 Ghent
Belgium
E-mail address: Haeike.Josephy@UGent.be

Tom Loeys
Department of Data Analysis
Ghent University
Henri Dunantlaan 2
9000 Ghent
Belgium
E-mail address: Tom.Loeys@UGent.be

Sara Kindt
Department of Experimental Clinical
and Health Psychology
Ghent University
Henri Dunantlaan 2
9000 Ghent
Belgium
E-mail address: Sara.Kindt@UGent.be