Bayesian joint modeling for causal mediation analysis with a binary outcome and a binary mediator: exploring the role of obesity in the association between cranial radiation therapy for childhood acute lymphoblastic leukemia treatment and the long-term risk of insulin resistance

Caubet, Miguel; Samoilenko, Mariia; Drouin, Simon; Sinnett, Daniel; Krajinovic, Maja; Laverdière, Caroline; Marcil, Valérie; Lefebvre, Geneviève

Publisher's version / Version de l'éditeur:
https://doi.org/10.1016/j.csda.2022.107586
Computational Statistics & Data Analysis, 177, C, 2022-08-08

Computational Statistics and Data Analysis 177 (2023) 107586

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Bayesian joint modeling for causal mediation analysis with a binary outcome and a binary mediator: Exploring the role of obesity in the association between cranial radiation therapy for childhood acute lymphoblastic leukemia treatment and the long-term risk of insulin resistance ☆

Miguel Caubet [a], Mariia Samoilenko [a], Simon Drouin [b,c], Daniel Sinnett [b,d], Maja Krajinovic [b,e], Caroline Laverdière [b,f], Valérie Marcil [b,g], Geneviève Lefebvre [a,*]

[a] Department of Mathematics, University of Quebec at Montreal (UQAM), 201 President-Kennedy Avenue, Montreal, QC, H2X 3Y7, Canada
[b] Division of Hematology-Oncology, Research Center, Sainte-Justine University Health Center, 3175 Chemin de la Côte-Sainte-Catherine, Montréal, QC, H3T 1C5, Canada
[c] Human Health Therapeutics Research Center, National Research Council Canada, 6100 Royalmount Avenue, Montréal, QC, H4P 2R2, Canada
[d] Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, QC, H3T 1C5, Canada
[e] Department of Pharmacology and Physiology, Faculty of Medicine, University of Montreal, Montreal, QC, H3T 1J4, Canada
[f] Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, QC, H3T 1A8, Canada
[g] Department of Nutrition, Faculty of Medicine, University of Montreal, Montreal, QC, H3T 1A8, Canada

ARTICLE INFO

Article history:
Received 2 October 2021
Received in revised form 15 April 2022
Accepted 29 July 2022
Available online 8 August 2022

Keywords:
Causal mediation analysis
Binary outcome
Binary mediator
Bayesian modeling
Multivariate logistic regression
Sensitivity analysis

ABSTRACT

Mediation analysis with a binary outcome is notoriously more challenging than with a continuous outcome. A new Bayesian approach for performing causal mediation with a binary outcome and a binary mediator, named the *t*-link approach, is introduced. This approach relies on the Bayesian multivariate logistic regression model introduced by O'Brien and Dunson (2004) and its Student-*t* approximation. By re-expressing the Mediation Formula, it is shown how to use this multivariate latent model for estimating the natural direct and indirect effects of an exposure on an outcome in any measure scale of interest (e.g., odds or risk ratio, risk difference). The *t*-link mediation approach has several valuable features which, to our knowledge, are not found together in existing binary-binary mediation analysis approaches. In particular, it allows for sensitivity analyses regarding the impact of unmeasured mediator-outcome confounders on the natural effects estimates. The proposed mediation approach was evaluated and compared with two other benchmark approaches using simulated data. Results revealed the usefulness of the *t*-link mediation approach when the sample size is small or moderate. Lastly, the *t*-link approach was applied for assessing the impact of cranial radiation therapy given to treat childhood acute lymphoblastic leukemia on the long-term risk of insulin resistance, where this effect is possibly mediated by obesity.

☆ R code programs for implementing the *t*-link mediation approach are available as electronic supplements.

* Corresponding author.
E-mail address: lefebvre.gen@uqam.ca (G. Lefebvre).

https://doi.org/10.1016/j.csda.2022.107586

1. Introduction

Mediation analyses are a modern analytical tool that has become ubiquitous for practitioners in many areas of applied research. Such analyses attempt to evaluate the association between an exposure A and an outcome Y of interest under the premise there exists an intermediate variable (mediator) M on the causal pathway between them. More precisely, the aim is to decompose the total effect of the exposure A on the outcome Y into two mutually complementary effects: the indirect effect of A on Y via M, and the remaining effect that is taken to be the direct effect. The psychological and biomedical fields have much contributed to the traditional and causal mediation analysis literatures, but a large fraction of advances has concerned continuous outcome variables. In fact, it is well known that the development of mediation approaches is more challenging when the outcome is binary, as opposed to continuous, due to the consideration of nonlinear models (Pearl, 2012; Loeys et al., 2013; Rijnhart et al., 2021). As binary variables are frequently encountered in practice, and are notably predominant in medical research, this motivates new methodological approaches for dealing with such a type of variables.

Mediation analysis approaches usually involve estimating regression coefficients from two or three separate models (e.g., Imai et al. (2010b); Lange et al. (2012); Tchetgen and Shpitser (2012); Valeri and VanderWeele (2013); Rijnhart et al. (2017)). For instance, in a conventional regression-based approach, we typically specify an outcome model $Y|M, A, \boldsymbol{C}$ and a mediator model $M|A, \boldsymbol{C}$, where $\boldsymbol{C}$ is a set of adjustment covariates. These models are then fitted separately to data and results combined to yield direct and indirect effects estimates. However, we could envisage performing a mediation analysis directly from a bivariate model $Y, M|A, \boldsymbol{C}$, where the estimation is done in one single step. To our knowledge, joint perspectives to mediation are a relatively new and underexplored idea in the literature (Enders et al., 2013; Albert and Wang, 2015; Wagner et al., 2018; Wang et al., 2019). In the Bayesian framework, joint modeling with data augmentation for missing data was introduced by Enders et al. (2013), but only for a continuous mediator and a continuous outcome through a multivariate normal model for (Y, M, A). In the context of sensitivity analyses for unmeasured mediator-outcome confounding, Albert and Wang (2015) proposed a joint model suitable for responses following generalized linear models in the form of a Gaussian copula model involving (possibly latent) mediator and outcome variables. A joint perspective to mediation was also advocated by Wagner et al. (2018), who introduced a multivariate generalized linear model approach with mixed variable types to facilitate hypothesis testing and construction of confidence ellipses for the indirect effect. Their approach targets the joint distribution $Y, M|A, \boldsymbol{C}$, but nonetheless involves specifying an outcome and a mediator model separately (that is, models $Y|M, A, \boldsymbol{C}$ and $M|A, \boldsymbol{C}$), where the two models' log likelihoods are summed up to determine the bivariate model log likelihood.

In this article, we develop a joint perspective to causal mediation analysis in the Bayesian parametric framework while specifically targeting a binary outcome and a binary mediator. Our proposed method, referred herein as the t-link mediation analysis approach, relies on the multivariate logistic regression model introduced by O'Brien and Dunson (2004) and its Student-t (t-link) approximation (O'Brien and Dunson, 2004; Hund et al., 2015). Concretely, our approach requires utilizing the multivariate t-link model to calculate the expected nested counterfactual outcomes that underlie the definitions of the natural direct and indirect effects (NDE and NIE, respectively). The model is fitted using Markov chain Monte Carlo (MCMC) and inference on the NDE and NIE is performed using posterior draws of the model's parameters.

Our proposed approach falls in line with the growing interest in the development of Bayesian methods for mediation analysis since the last fifteen years (Yuan and MacKinnon, 2009; Elliott et al., 2010; Daniels et al., 2012; Park and Kaplan, 2015; Kim et al., 2017, 2018; Miočević et al., 2018; Song et al., 2020, 2021). There are several reasons to favor Bayesian methods in mediation. First, it is possible to incorporate relevant prior information, which may yield increased statistical power (Miočević et al., 2017). Second, the inference is deemed exact in that it does not rely on large-sample approximations based on normality, which can be inappropriate when the sample size is small (Yuan and MacKinnon, 2009). This is particularly useful for characterizing mediation, as analyses are prone to inaccuracies if the asymmetric distribution of the indirect effect estimator is not acknowledged in inference (MacKinnon et al., 2004; Yuan and MacKinnon, 2009). Third, Bayesian methods allow for a probabilistic interpretation of results, as discussed by Miočević et al. (2018) in the context of mediation analyses. For the specific case of a binary outcome and a binary mediator, Elliott et al. (2010) proposed Bayesian principal stratification for estimating direct and indirect effects, while assuming the exposure is assigned at random. Kim et al. (2017) developed a Bayesian nonparametric approach based on Dirichlet process mixtures for estimating natural effects with continuous or binary outcome and mediator variables, but the approach was assessed and illustrated with a continuous outcome and a continuous mediator only.

The t-link mediation approach possesses valuable features which, to our knowledge, are not found together in current causal or traditional binary-binary mediation approaches. To begin, the bivariate specification for the latent variables underlying the t-link model is based on an approximate logistic distribution rather than a normal distribution, which implies a familiar logistic regression parameterization for $Y|A, \boldsymbol{C}$ and $M|A, \boldsymbol{C}$. Other interesting features originate from the Bayesian nature of our proposed approach, as described previously. In particular, since the t-link approach is implemented using MCMC, mediation effects and proportions mediated are easily reported on all standard binary scales (e.g., risk difference, risk ratio, odds ratio). Further, this approach is valid independently of the rareness or commonness of the outcome, which is worthwhile since there has been a recent awareness that the rare outcome assumption is difficult to formalize and assess in practice in mediation set-ups (Samoilenko and Lefebvre, 2018; VanderWeele et al., 2018; Samoilenko and Lefebvre, 2021).

One last important feature is the fact that the t-link mediation approach easily allows for sensitivity analyses regarding the impact of unmeasured mediator-outcome confounders on the natural effects estimates.

Our paper is divided as follows. In Section 2, we review the definitions of the NDE and NIE and introduce the multivariate logistic model proposed by O'Brien and Dunson (2004) and its t-link approximation. We then explain how the t-link model can be applied for mediation analyses and used to estimate natural effects. In this section, we also discuss the t-link representation of the total exposure effect on the odds ratio scale. In Section 3, we present a simulation study investigating the performance of the t-link approach for mediation in scenarios corresponding to various data generating mechanisms. More precisely, our mediation approach is compared to two other (semi)parametric approaches, namely the weighting-based natural effect model approach of Lange et al. (2012) and the recent exact regression-based approach for a binary outcome and a binary mediator by Samoilenko and Lefebvre (2021). In Section 4, we describe how a sensitivity analysis for unmeasured mediator-outcome confounding can be performed using the proposed model and provide an evaluation using simulated data. In Section 5, we apply the t-link mediation approach to get insight on the cardiometabolic health of childhood acute lymphoblastic leukemia (cALL) survivors. More precisely, the interest lies in the potential effect of cranial radiation therapy received to treat cALL on survivors' insulin resistance, where this effect is possibly mediated by obesity. Our cohort data come from the PETALE study (Marcoux et al., 2017), which was conducted at Sainte-Justine University Health Center between 2012-2016 to identify and characterize the most common late adverse effects observed among cALL survivors. The data, of moderate size ($n = 245$), are perfectly suited for this Bayesian binary-binary mediation analysis.

2. Methods

2.1. Natural effects for a binary outcome and a binary mediator

Natural direct and indirect effects are used in causal mediation analyses to determine how much of the effect of the exposure (treatment) on the outcome naturally goes through the mediator, and how much goes through other mechanisms. Natural effects are defined on the basis of the nested counterfactual outcome $Y(a, M(a^*))$, which is the outcome that would have been observed if exposure A had been set to a and mediator M had been set to the value it would have taken if A had been set to a^*, where $a, a^* \in \mathcal{A}$ and $\mathcal{A}$ is the set of possible values for A (Robin and Greenland, 1992; Pearl, 2001). When $a^* = a$, this nested counterfactual outcome is taken to correspond to the counterfactual outcome one would observe if A had been set to a: $Y(a, M(a)) = Y(a)$ (VanderWeele and Vansteelandt, 2009).

When the outcome is binary, the conditional NDE and NIE, comparing exposure levels a and a^* where a^* indicates a reference or baseline exposure value different than a (i.e. $a \neq a^*$), are often expressed on the odds ratio scale:

$$OR^{NDE}_{a,a^*|\boldsymbol{c}} = \frac{P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(a^*, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(a^*, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c}))},$$

$$OR^{NIE}_{a,a^*|\boldsymbol{c}} = \frac{P(Y(a, M(a)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(a, M(a)) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c}))},$$

where the total effect (TE) odds ratio is given by

$$OR^{TE}_{a,a^*|\boldsymbol{c}} = OR^{NDE}_{a,a^*|\boldsymbol{c}} \times OR^{NIE}_{a,a^*|\boldsymbol{c}}. \tag{1}$$

These natural effects can be similarly formulated on the risk ratio or risk difference scales.

Let $Y(a, m)$ be the counterfactual outcome if A had been set to a and M had been set to m. Natural direct and indirect effects can be identified from data under several strong assumptions. These assumptions first include the consistency of the counterfactuals related to the outcome and mediator ($Y = Y(a)$ if $A = a$; $Y = Y(a, m)$ if $A = a$ and $M = m$; $M = M(a)$ if $A = a$), composition ($Y(a, M(a)) = Y(a)$), no interference (i.e., the counterfactuals related to the outcome and mediator for a given individual are not affected by the exposure level of another individual), and positivity ($P(A = a|\boldsymbol{C} = \boldsymbol{c}) > 0 \ \forall a \in \mathcal{A}$; $P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) > 0 \ \forall m \in \mathcal{M}$, where $\mathcal{M}$ is the set of possible values for M) (VanderWeele and Vansteelandt, 2009; Lindmark et al., 2018; Nguyen et al., 2020). Then unconfoundedness identifiability assumptions are (VanderWeele and Vansteelandt, 2009; Nguyen et al., 2020):

A_1: no-unmeasured confounders for the exposure-outcome relationship ($Y(a, m) \perp\!\!\!\perp A|\boldsymbol{C} = \boldsymbol{c} \quad \forall a \in \mathcal{A}$ and $\forall m \in \mathcal{M}$);
A_2: no-unmeasured confounders for the mediator-outcome relationship ($Y(a, m) \perp\!\!\!\perp M|A = a, \boldsymbol{C} = \boldsymbol{c} \quad \forall a \in \mathcal{A}$ and $\forall m \in \mathcal{M}$);
A_3: no-unmeasured confounders for the exposure-mediator relationship ($M(a) \perp\!\!\!\perp A|\boldsymbol{C} = \boldsymbol{c} \quad \forall a \in \mathcal{A}$);
A_4: cross-world assumption, which is also often presented as no mediator-outcome confounders affected by exposure ($Y(a, m) \perp\!\!\!\perp M(a^*)|\boldsymbol{C} = \boldsymbol{c} \ \ \forall a, a^* \in \mathcal{A}$ and $\forall m \in \mathcal{M}$).

Under above assumptions, the nested counterfactual outcome probabilities can be written in terms of probabilities defined on observable variables. For a binary mediator M, the *Mediation Formula* (Pearl, 2001) links these probabilities as follows:

$$P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c}) = \sum_{m=0,1} P(Y = 1|M = m, A = a, \boldsymbol{C} = \boldsymbol{c}) P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c}). \tag{2}$$

Each nested counterfactual outcome probability involved in the definitions of the NDE and NIE can then be estimated by estimating each term in the sum in (2).

2.2. t-link model for binary-binary mediation

Computational and modeling reasons have contributed to the widespread use of probit models for modeling multivariate correlated binary responses, especially in the Bayesian framework (O'Brien and Dunson, 2004). While probit models are attractive since they permit to easily model the latent normal variables' dependency structure in terms of correlation coefficients, they are known for lacking straightforward interpretation of corresponding regression coefficients (O'Brien and Dunson, 2004). Grounded in the Bayesian framework, the multivariate logistic regression model introduced by O'Brien and Dunson (2004) is an appealing model to study the association between a set of covariates and several correlated binary response variables. It is based on the observation that a p-dimensional multivariate logistic distribution can be induced by transforming variables that follow a conventional multivariate distribution (e.g., multivariate normal or multivariate Student-t distribution) in such a way that transformed variables have a logistic distribution marginally; O'Brien and Dunson (2004) proposed to use Student-t distributed variables with a scale matrix $\mathbf{R}$ restricted to have 1's on the diagonal but otherwise unrestricted. This model then yields results with high interpretability, by allowing for the familiar log odds ratio interpretation of regression coefficients as in a standard univariate logistic model, while permitting flexible dependency structures between response variables.

As explained in O'Brien and Dunson (2004), an incentive for selecting a multivariate Student-t distribution for the underlying variables is the fact that the corresponding multivariate logistic model can be approximated by a multivariate Student-t model (t-link model), which greatly facilitates the implementation of the MCMC algorithm. The authors motivated the approximation by the fact that a univariate Student-t distribution, chosen with specific scale parameter and number of degrees of freedom, closely resembles the univariate logistic distribution. In O'Brien and Dunson (2004), the regression coefficients estimates obtained from the t-link model were found to have practically no differences with those pertaining to the logistic model. Hence we adopt herein the Student-t representation of the multivariate logistic model, which approximately retains the log odds ratio interpretation of coefficients while being significantly easier from a computational point of view. This Student-t representation was also retained by Hund et al. (2015) for estimating average treatment effects.

In the sequel, we present the t-link model for binary-binary mediation and show, in Section 2.2.2, how it can be used to provide a parametric formulation of the NDE and NIE.

2.2.1. Model specification

A p-dimensional vector $\boldsymbol{T}$ follows a multivariate Student-t distribution with ν degrees of freedom, location vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$ if its density is given by

$$St(\boldsymbol{t};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left(1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right)^{-(p+\nu)/2}. \tag{3}$$

Let $\boldsymbol{Z}_i=(Z_i^Y,Z_i^M)'$ be the binary (1/0) vector of responses for individual i, $i=1,\dots,n$, where Y and M index the outcome and mediator, respectively. We define $\boldsymbol{L}_i=(L_i^Y,L_i^M)'$ as a bivariate vector of Student-t latent variables with mean $\boldsymbol{\mu}_i=(\mu_i^Y,\mu_i^M)'$ where

$$\mu_i^j=\beta_0^j+\beta_A^jA_i+\boldsymbol{\beta}_{\boldsymbol{I}}^{j'}A_i\boldsymbol{C}_i+\boldsymbol{\beta}_{\boldsymbol{C}}^{j'}\boldsymbol{C}_i,\quad j\in\{Y,M\}, \tag{4}$$

and A_i and $\boldsymbol{C}_i$ are the exposure and covariates for individual i, respectively, and the $\boldsymbol{\beta}^j$s are unknown regression coefficients to be estimated. As in O'Brien and Dunson (2004), we take the scale matrix as $\boldsymbol{\Sigma}=\delta^2\mathbf{R}$, where $\mathbf{R}$ is a correlation matrix, $\delta^2=\pi^2\frac{(\nu-2)}{(3\nu)}$, and $\nu=7.3$. Because only two response variables are involved in the model (the mediator and the outcome), $\mathbf{R}$ is of dimension 2×2 and features a single unknown correlation coefficient ρ. Because interaction is often allowed in mediation models (Muller et al., 2005; Valeri and VanderWeele, 2013), we include interaction terms between exposure and covariates in the mean models (4) for L_i^Y and L_i^M. Finally, the outcome and mediator variables are linked to the latent variables by $Z_i^j=I(L_i^j>0)$, where $I(\cdot)$ is the indicator function and $j\in\{Y,M\}$. Henceforth we take $Y_i\equiv Z_i^Y$ and $M_i\equiv Z_i^M$.

The proposed t-link model allows for calculating the conditional probability of a given configuration of the response vector under each level of exposure ($A=a,a^*$):

$$P(Y=y,M=m|A=a,\boldsymbol{C}=\boldsymbol{c})\quad\text{and}\quad P(Y=y,M=m|A=a^*,\boldsymbol{C}=\boldsymbol{c}), \tag{5}$$

where $m,y\in\{0,1\}$. For each of the $2^2=4$ possible outcome and mediator configurations, probabilities in (5) are obtained by integrating the latent variables bivariate Student-t density function over intervals compatible with the configuration. For example,

$$P(Y=1,M=0|A=a,\boldsymbol{C}=\boldsymbol{c})=P(L^Y>0,L^M\le 0|A=a,\boldsymbol{C}=\boldsymbol{c}),$$

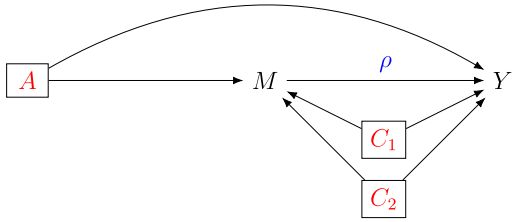

Fig. 1. A directed acyclic graph for mediation in which covariates C_1 and C_2 are confounders for the $M-Y$ relationship. The direct arrow from M to Y is the sole contributor to the magnitude of the residual dependency between M and Y (encoded in ρ) in a t-link model that adjusts for C_1 and C_2.

$$= \int_{0}^{\infty} \int_{-\infty}^{0} St(l^Y, l^M; \mu^Y, \mu^M, \delta^2\mathbf{R}, \nu) dl^M dl^Y, \tag{6}$$

where $\mu^Y = \beta_0^Y + \beta_A^Y a + \boldsymbol{\beta}_{\boldsymbol{I}}^{Y'} a\boldsymbol{c} + \boldsymbol{\beta}_{\boldsymbol{C}}^{Y'} \boldsymbol{c}$ and $\mu^M = \beta_0^M + \beta_A^M a + \boldsymbol{\beta}_{\boldsymbol{I}}^{M'} a\boldsymbol{c} + \boldsymbol{\beta}_{\boldsymbol{C}}^{M'} \boldsymbol{c}$.

In the proposed mediation model, the dependency between the mediator M and outcome Y after accounting for A and $\boldsymbol{C}$ is encoded in the off-diagonal element ρ of $\mathbf{R}$. The association induced by the presence of an arrow from M to Y thus determines ρ through latent variables L^Y and L^M. If not adjusted for in the model, any covariate that is a common cause of M and Y also contributes to the association between M and Y and therefore impacts on ρ. However, in causal mediation, we require that all confounders for the $M-Y$ relationship be measured and adjusted for (recall A_2) and, as such, these covariates are not expected to contribute to the residual dependency between M and Y. Thus, in the hypothesized setting, ρ solely reflects the possible direct link from M to Y necessary for a mediated effect of the exposure on the outcome. Here we also eliminate the possibility that an association between M and Y is induced by inadvertently adjusting on a common effect of M and Y (i.e., collider) (Cole et al., 2009). We refer the reader to Fig. 1 for a visual interpretation of ρ under the hypothesis of no unmeasured confounders between the mediator and outcome.

2.2.2. Mediation formula with joint outcome and mediator probabilities

The usual presentation of the Mediation Formula involves a conditional expectation (probability) for the outcome given the mediator as shown in (2). For our purpose, we instead write the formula in terms of a joint outcome and mediator probability as follows:

$$\begin{aligned} P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c}) &= \sum_m P(Y = 1|M = m, A = a, \boldsymbol{C} = \boldsymbol{c}) P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c}) \\ &= \sum_m P(Y = 1, M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) \frac{P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c})}{P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c})}, \end{aligned} \tag{7}$$

where $P(M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) = \sum_y P(Y = y, M = m|A = a, \boldsymbol{C} = \boldsymbol{c})$ and similarly for $P(M = m|A = a^*, \boldsymbol{C} = \boldsymbol{c})$. Whenever $a^* = a$, the nested counterfactual outcome probability $P(Y(a, M(a^*)) = 1|\boldsymbol{C} = \boldsymbol{c})$ can be simplified as $P(Y(a, M(a)) = 1|\boldsymbol{C}) = \sum_m P(Y = 1, M = m|A = a, \boldsymbol{C} = \boldsymbol{c}) = P(Y = 1|A = a, \boldsymbol{C} = \boldsymbol{c})$.

For a binary exposure with levels $a, a^* \in \{0, 1\}$, the NDE involves probabilities $P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})$ and $P(Y(0, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})$, for each measure scale selected. Similarly, the NIE requires $P(Y(1, M(1)) = 1|\boldsymbol{C} = \boldsymbol{c})$ and $P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})$. Each of these nested counterfactual outcome probabilities can be written on the basis of (7) using joint probabilities expressed as in (6), thereby providing the t-link model-based definitions for the natural effects.

2.2.3. t-link representation of the total exposure effect on the odds ratio scale

In the standard regression-based causal mediation approach for a binary outcome and a binary mediator, two logistic regression models are specified, but the logistic model for the outcome Y is expressed conditionally on the mediator M (e.g., Valeri and VanderWeele (2013); Samoilenko and Lefebvre (2021)). The implied marginal model $Y|A, \boldsymbol{C}$ is therefore not logistic (Wang, 2020). Opposingly, the t-link mediation approach fixes both marginal models to be (approximately) logistic while the conditional model for Y given $M, A, \boldsymbol{C}$ follows implicitly. We believe that the marginal logistic specification for the outcome is a nice feature of the proposed t-link approach since it perfectly matches the habitual choice of a logistic regression model for $Y|A, \boldsymbol{C}$ when performing a classical (non-mediated) analysis for exposure effect estimation. As is discussed below, our mediation model thus allows to parcimoniously encode, in a single coefficient, the total effect of the exposure on the binary outcome on the odds ratio scale.

Consider the O'Brien and Dunson (2004) multivariate logistic regression model with an arbitrary number of binary response variables, with outcome Y being one of them. Suppose that the model is correctly specified and involves binary exposure variable A and covariates $\boldsymbol{C}$ as regressors. Covariates may be included in different functional forms in the model, but we assume there are no interaction terms between A and $\boldsymbol{C}$. Owing to the logistic interpretation of the model, the coefficient associated with the exposure in the marginal model for Y (β_A^Y) can be interpreted as the log of the conditional odds

ratio for the outcome for exposed ($A = 1$) versus unexposed ($A = 0$). If the set of covariates $\boldsymbol{C}$ is sufficient to control for confounding and that other assumptions (no interference and positivity) are satisfied, then the coefficient β_A^Y has a causal interpretation and represents the effect of exposure A on the outcome Y. Indeed, as observed by Hund et al. (2015), the (non-mediated) exposure effect of A on Y on the odds ratio scale can be written as:

$$OR_{1,0|\boldsymbol{c}} = \frac{P(Y(1) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(0) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(0) = 1|\boldsymbol{C} = \boldsymbol{c}))} = exp(\beta_A^Y).$$

Now consider this model for mediation analysis, such that one of the two responses is the outcome Y and the other is the mediator M, as proposed previously. Then the total exposure effect $OR_{1,0|\boldsymbol{c}}^{TE}$ can be written as:

$$\begin{aligned}
OR_{1,0|\boldsymbol{c}}^{TE} &= OR_{1,0|\boldsymbol{c}}^{NDE} \times OR_{1,0|\boldsymbol{c}}^{NIE} \\
&= \frac{P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(0, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(0, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c}))} \\
&\quad \times \frac{P(Y(1, M(1)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1, M(1)) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c}))} \\
&= \frac{P(Y(1, M(1)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1, M(1)) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(0, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(0, M(0)) = 1|\boldsymbol{C} = \boldsymbol{c}))} \qquad (8) \\
&= \frac{P(Y(1) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(1) = 1|\boldsymbol{C} = \boldsymbol{c}))}{P(Y(0) = 1|\boldsymbol{C} = \boldsymbol{c})/(1 - P(Y(0) = 1|\boldsymbol{C} = \boldsymbol{c}))} \qquad (9) \\
&= OR_{1,0|\boldsymbol{c}} = exp(\beta_A^Y),
\end{aligned}$$

where the step from (8) to (9) holds by the composition assumption and where the model-based definitions for $OR_{1,0|\boldsymbol{c}}^{NDE}$ and $OR_{1,0|\boldsymbol{c}}^{NIE}$ can be written as explained in Section 2.2.2. Put differently, the product of the NDE and NIE parametrized from this model coincides, on the odds ratio scale, with the exponential of coefficient β_A^Y.

3. Simulation study

We conducted a simulation study to evaluate the performance of the t-link model for mediation in a variety of scenarios. Our study first considered three standard scenarios, each featuring a non null indirect effect (*Main Scenarios*). Then, we studied the appropriateness of the t-link model in a scenario in which the outcome and mediator were independent given exposure and covariates and the indirect effect was null (*No Mediation Scenario*). Next, we examined four mediation scenarios in which the data generating mechanism included an interaction term involving the exposure variable (*Interaction Scenarios*). For these scenarios, two of them were compatible with the standard counterfactual specification which considers an exposure-mediator interaction term in the outcome model. All simulation scenarios were selected to illustrate and highlight some aspects in the performance of the t-link approach proposed.

3.1. Main mediation scenarios

In the first scenario (*Main Scenario 1*), the covariate C_1 was generated as a Bernoulli variable with probability $p = 0.5$ ($C_1 \sim Ber(p = 0.5)$), whereas covariate C_2 was generated independently according to a standard normal variable ($C_2 \sim N(0, 1)$). The exposure A was generated as a Bernoulli variable with probability $p_{\boldsymbol{c}} = expit(\log(0.7) + \log(2)c_1 + \log(1.5)c_2)$, where $expit(\cdot) = exp(\cdot)/(1 + exp(\cdot))$. The binary outcome and mediator were then generated according to the O'Brien and Dunson (2004) logistic regression model with correlation fixed to $\rho = 0.7$ and with mean $\boldsymbol{\mu} = (\mu^Y, \mu^M)'$, where

$$\begin{aligned}
\mu^Y &= -\log(2) + \log(2.8)A + \log(3)C_1 - \log(1.5)C_2, \\
\mu^M &= -\log(2) + \log(4.5)A + \log(3)C_1 - \log(2)C_2.
\end{aligned}$$

We refer the reader to Appendix A.1 for more details on the simulation of data.

In the second scenario (*Main Scenario 2*), we considered two additional independent covariates, C_3 and C_4, in addition to those considered in *Main Scenario 1* ($C_3 \sim Ber(p = 0.3)$, $C_4 \sim N(0, 1)$). The exposure A was generated as a Bernoulli variable with probability $p_{\boldsymbol{c}} = expit(\log(0.8) + \log(2)c_1 + \log(1.5)c_2 - \log(1.4)c_3 - \log(1.2)c_4)$. Data were generated using the O'Brien and Dunson (2004) logistic regression model with $\rho = 0.2$ and the following parameter values:

$$\begin{aligned}
\boldsymbol{\beta}^Y &= (\beta_0^Y, \beta_A^Y, \beta_{C_1}^Y, \beta_{C_2}^Y, \beta_{C_3}^Y, \beta_{C_4}^Y)' \\
&= (\log(0.8), \log(3), \log(2), -\log(1.5), \log(2), -\log(1.7))', \\
\boldsymbol{\beta}^M &= (\beta_0^M, \beta_A^M, \beta_{C_1}^M, \beta_{C_2}^M, \beta_{C_3}^M, \beta_{C_4}^M)' \\
&= (\log(1.3), \log(3), -\log(2), 0, -\log(1.4), \log(1.7))'.
\end{aligned}$$

The third scenario (*Main Scenario 3*) involved the same four covariates as in *Main Scenario 2*, with the model parameters values modified as $\rho = 0.7$ and

$$\begin{aligned}\boldsymbol{\beta}^Y &= (\beta_0^Y, \beta_A^Y, \beta_{C_1}^Y, \beta_{C_2}^Y, \beta_{C_3}^Y, \beta_{C_4}^Y)' \\ &= (-\log(2), \log(2.8), \log(3), -\log(1.5), -\log(1.6), \log(1.7))', \\ \boldsymbol{\beta}^M &= (\beta_0^M, \beta_A^M, \beta_{C_1}^M, \beta_{C_2}^M, \beta_{C_3}^M, \beta_{C_4}^M)' \\ &= (-\log(2), \log(4.5), \log(3), -\log(2), \log(2), -\log(2))'.\end{aligned}$$

For all scenarios investigated, 2000 datasets with three sample sizes ($n = 100, 250, 1000$ for *Main Scenario 1*; $n = 250, 500, 1500$ for *Main Scenarios 2-3*) were generated according to the corresponding data generating mechanism. Our rationale for the values of the sample size n is that Bayesian analyses are often indicated and used when samples are small. As such, it is relevant to evaluate proposed t-link model in this context. As for the Monte Carlo sample size (that is, the number of datasets), we selected 2000 to balance computational burden, since each dataset itself requires MCMC simulation, and stable estimation of performance metrics considered (bias, standard deviation, root mean squared error, coverage probability and mean interval length).

We fitted the t-link model for mediation to each dataset generated. The prior distribution was specified as $\pi(\boldsymbol{\beta}, \mathbf{R}) = \pi(\boldsymbol{\beta})\pi(\mathbf{R})$, where $\pi(\boldsymbol{\beta})$ is multivariate normal and $\pi(\mathbf{R})$ is uniform on the space of correlation matrices. More precisely, $\pi(\boldsymbol{\beta}) = \prod_j N(\boldsymbol{\beta}^j; \boldsymbol{\beta}_*^j, \boldsymbol{\Sigma}_*^j)$, where $\boldsymbol{\beta}_*^j = \mathbf{0}$ and $\boldsymbol{\Sigma}_*^j = diag(1000, 4, \ldots, 4)$ for $j \in \{Y, M\}$. Next, we sampled 110 000 draws from the posterior distribution of $(\boldsymbol{\beta}, \mathbf{R})$ and discarded the first 10 000 as burn-in. To alleviate chain post-processing while reducing autocorrelation between successive draws for the off-diagonal element of $\mathbf{R}$ (ρ), a thinning with a factor of 10 was applied, leaving a total of 10 000 draws available for inference. For each retained posterior draw for $(\boldsymbol{\beta}, \mathbf{R})$, nested counterfactual probabilities $P(Y(1, M(1)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})$, $P(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})$ and $P(Y(0, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})$ were calculated as described in Section 2.2.2. These conditional probabilities were computed with the value of each covariate set at its mean value in the sample ($\bar{\boldsymbol{c}}$). Finally, posterior draws of the NDE and NIE (measured on the OR scale), as well as corresponding proportion mediated, were computed from these counterfactual probabilities. Because the posterior distributions of ORs were skewed, the posterior medians were used as point estimators of the true corresponding conditional natural effects. Equal-tailed 95% credible intervals were obtained for the natural effects and proportion mediated. For *Main Scenario 1*, the NDE and NIE were also estimated on the RR and RD scales for illustration purposes. For the RR scale, the posterior median was used as the point estimator while the posterior mean was used for the RD scale. More details on the computation of natural effects are provided in Appendix A.2.

For each dataset generated, we also obtained conditional natural effects estimates using the weighting-based natural effect model (NEM) approach of Lange et al. (2012), implemented in the R package `medflex` (Steen et al., 2017). This approach requires weighting an expanded dataset using a ratio-mediator-probability weight. For a binary exposure, the expansion is done by duplicating each individual observation and creating an additional hypothetical exposure variable A^* that takes the values A for one replication and $1 - A$ for the other. The ratio-mediator-probability weight $P(M = M_i|A = A_i^*, \boldsymbol{C} = \boldsymbol{C}_i)/P(M = M_i|A = A_i, \boldsymbol{C} = \boldsymbol{C}_i)$ is then calculated for each individual i in the expanded dataset. One can note that the same mediator ratio is involved in our alternative presentation of the Mediation Formula shown in (7). Lastly a suitable mean model for the nested counterfactual outcome is specified and then fitted by using corresponding weighted model for the outcome Y. In our application of the NEM, we used a logistic model for the mediator to calculate the weights, where the functional form of the model matched the generating mediator model for a given scenario. The following logistic model was considered to parametrize the natural effects:

$$\text{logit}\, \mathrm{P}(Y(a, M(a^*)) = 1|\boldsymbol{C}) = \alpha_0 + \alpha_1 a + \alpha_2 a^* + \boldsymbol{\alpha}_3' \boldsymbol{C}. \tag{10}$$

The NDE and NIE on the odds ratio scale are given by the exponential of parameters α_1 and α_2, respectively, thereby revealing that conditional natural effects derived from (10) are the same for any level $\boldsymbol{c}$ of the adjustment covariates. For the NDE for instance,

$$\mathrm{OR}_{1,0|\boldsymbol{C}}^{NDE} = \frac{\text{odds}[P(Y(0+1, M(0)) = 1|\boldsymbol{C})]}{\text{odds}[P(Y(0, M(0)) = 1|\boldsymbol{C})]} = \exp(\alpha_1).$$

Natural effects estimates are obtained by replacing α_1 and α_2 by their estimates. Wald-type confidence intervals (based on robust standard errors) as well as percentile bootstrap confidence intervals (with 10 000 bootstrap replications) were obtained.

A number of works have recently focused on the development of exact regression-based approaches for the estimation of natural effects for a binary outcome and a binary mediator (Samoilenko et al., 2018; Gaynor et al., 2019; Samoilenko and Lefebvre, 2021; Doretti et al., 2022). The Exact mediation approach for a binary outcome and a binary mediator by Samoilenko and Lefebvre (2021), newly implemented in the R package `ExactMed` (Caubet et al., 2022), was also considered in our simulation study. This approach involves specifying a logistic model for Y given $M, A, \boldsymbol{C}$ and a logistic model for M given $A, \boldsymbol{C}$. Closed-form formulas for the NDE and NIE are then obtained on the basis of the Mediation Formula (2). This

exact regression-based approach is similar to the Valeri and VanderWeele (2013) approach, but the former does not make the simplifying assumption of a rare outcome. In our simulations, we implemented this approach using an outcome model with no exposure-mediator interaction term. We obtained exact conditional natural effects estimates on the OR scale and both delta and percentile bootstrap confidence intervals are reported.

3.2. No mediation scenario

We also investigated the appropriateness of the t-link model for mediation in a situation wherein the mediator is not a direct cause of the outcome, and therefore the indirect effect of the exposure on the outcome is null. One caveat of the t-link model stems from the fact that a null correlation between the latent mediator and outcome variables does not imply that these latent variables are independent. Indeed, it is relatively well known that the multivariate Student-t density (3) cannot be expressed as the product of independent univariate Student-t densities when $\rho = 0$. As this feature of the t-link model was left unmentioned in O'Brien and Dunson (2004), we believe important to highlight the incongeniality of this model to independence between response variables.

To investigate this issue in our mediation context, we considered the following generating models (*No Mediation Scenario*):

$$\begin{aligned}&\operatorname{logit} \mathrm{P}(Y=1|A=a, M=m, C_1=c_1, C_2=c_2) = \\ &\qquad -\log(2)+\log(2.8)a+0m+\log(3)c_1-\log(1.5)c_2, \end{aligned} \tag{11}$$

$$\operatorname{logit} \mathrm{P}(M=1|A=a, C_1=c_1, C_2=c_2) = -\log(2)+\log(4.5)a+\log(3)c_1-\log(2)c_2. \tag{12}$$

This data generating mechanism is similar to *Main Scenario 1* since the outcome and mediator are each distributed according to a logistic model with the same coefficients. However, in the *No Mediation Scenario*, the mediator and outcome variables are independent given $A, \boldsymbol{C}$, unlike in *Main Scenario 1*.

For this scenario, the true values for the NIE were either 1 for the OR and RR scales or 0 for the RD scale. The corresponding true NDE values were obtained based on counterfactual probabilities computed using the Mediation Formula (2) with generating models (11) and (12). For the OR scale, the NDE coincides with the exponentiated coefficient associated with the exposure in the outcome model (11) (that is, $\text{NDE} = \exp(\beta_A^Y) = \exp(\log(2.8)) = 2.8$ in this scenario). We generated 2000 datasets of size $n = 100, 250, 1000$ under the *No Mediation Scenario*. The t-link model for mediation was then fitted to each dataset to estimate the natural effects on the OR, RR and RD scales, using the same approach as presented for the main simulation scenarios. Results for the NEM and Exact approaches, implemented as in the main scenarios, were also obtained for the OR scale.

3.3. Interaction scenarios

As for other causal mediation analysis approaches, the t-link mediation approach allows for the presence of interaction. In this section, we examine the performance of our approach to estimate natural effects within the context of data mechanisms featuring an exposure-covariate or an exposure-mediator interaction.

The first two interaction scenarios (*Interaction Scenarios 1 and 2*) involve data generated according to the studied model in which an interaction term of type "exposure-covariate" is included either in the mean model for M or the mean model for Y. More precisely, the outcome and mediator were generated according to the O'Brien and Dunson (2004) logistic regression model with mean $\boldsymbol{\mu} = (\mu^Y, \mu^M)'$, where

$$\mu^j = \beta_0^j + \beta_A^j A + \boldsymbol{\beta}_{\boldsymbol{I}}^{j'} A\boldsymbol{C} + \boldsymbol{\beta}_{\boldsymbol{C}}^{j'}\boldsymbol{C}, \quad j \in \{Y, M\}, \tag{13}$$

and residual correlation ρ.

In *Interaction Scenario 1*, we considered an $A - C_1$ interaction term I in the outcome model. This scenario allowed for so-called mediated moderation (Muller et al., 2005) in which there was a total effect modification by the baseline covariate C_1. The parameters values of the model were:

$$\begin{aligned}\boldsymbol{\beta}^Y &= (\beta_0^Y, \beta_A^Y, \beta_I^Y, \beta_{C_1}^Y, \beta_{C_2}^Y)' = (\log(0.8), \log(3), \log(1.8), \log(2), -\log(1.5))', \\ \boldsymbol{\beta}^M &= (\beta_0^M, \beta_A^M, \beta_{C_1}^M, \beta_{C_2}^M)' = (\log(1.3), \log(3), -\log(2), 0)',\end{aligned}$$

with $\rho = 0.5$, where covariates C_1 and C_2, and exposure A were generated the same way as in *Main Scenario 1*.

In *Interaction Scenario 2*, we instead considered an $A - C_1$ interaction term I in the mediator model. This scenario allowed for so-called moderated mediation (Muller et al., 2005) in which there was an indirect effect (and direct effect) modification by the baseline covariate C_1, while the total effect remained the same for either C_1 value (on the OR scale). The parameters values of the model were:

$$\begin{aligned}\boldsymbol{\beta}^Y &= (\beta_0^Y, \beta_A^Y, \beta_{C_1}^Y, \beta_{C_2}^Y)' = (\log(0.8), \log(3), \log(2), -\log(1.5))', \\ \boldsymbol{\beta}^M &= (\beta_0^M, \beta_A^M, \beta_I^M, \beta_{C_1}^M, \beta_{C_2}^M)' = (\log(1.3), \log(1.2), \log(2.2), -\log(2), 0)',\end{aligned}$$

with $\rho = 0.2$.

For the third and fourth interaction scenarios (*Interaction Scenarios 3 and 4*), we generated data using two separate logistic regression models, one for the mediator and one for the outcome given mediator, wherein an exposure-mediator interaction term was included in the outcome model:

$$\text{logit}\,\text{P}(Y = 1|A = a, M = m, C_1 = c_1, C_2 = c_2) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4 c_1 + \theta_5 c_2\,, \tag{14}$$

$$\text{logit}\,\text{P}(M = 1|A = a, C_1 = c_1, C_2 = c_2) = \gamma_0 + \gamma_1 a + \gamma_2 c_1 + \gamma_3 c_2\,. \tag{15}$$

Models presented in (14) and (15) constitute a well-known regression-based specification for a binary mediator and a binary outcome (e.g., VanderWeele (2015), Samoilenko and Lefebvre (2021)). For *Interaction Scenario 3*, the regression coefficient values were:

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)' = (-\log(2), \log(2.8), \log(1.5), -\log(2.8), \log(3), -\log(1.5))'\,,$$

$$\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)' = (-\log(2), \log(4.5), -\log(3), \log(5))'\,.$$

Interaction Scenario 4 was specified the same way as *Interaction Scenario 3* except for θ_3 which was equal to $-\log(1.68)$ instead of $-\log(2.8)$.

We generated 2000 datasets with three sample sizes ($n = 250, 500, 1500$) for each interaction scenario investigated. Natural effects estimates on the OR scale were obtained using the t-link model for mediation. For *Interaction Scenarios 1 and 2*, the fitted t-link model included the same terms as in the generating model, hence the working model was (approximately) correctly specified. For *Interaction Scenarios 3 and 4*, the fitted t-link model included either no interaction term or both $A - C_1$ and $A - C_2$ terms in the outcome mean model. In the latter case, these two interaction terms were included so to investigate whether they could account for the $A - M$ interaction present in the generating model (14). Indeed, because our proposed mediation model is a joint model that specifies a marginal model for M and Y separately, an interaction term $A - M$ cannot be included in the outcome mean model. However, since the mediator M is a function of covariates $\boldsymbol{C}$, the presence of an interaction term of type $A - M$ such as in (14) implies that the exposure and covariates interact together - to some extent - to explain the outcome.

Results for the NEM and Exact approaches were also obtained for these interaction scenarios. In *Interaction Scenario 1*, the structural model for the NEM was

$$\text{logit}\,\text{P}(Y(a, M(a^*)) = 1|C_1 = c_1, C_2 = c_2) = \alpha_0 + \alpha_1 a + \alpha_2 a^* + \alpha_3 a c_1 + \alpha_4 a^* c_1 + \alpha_5 c_1 + \alpha_6 c_2. \tag{16}$$

For *Interaction Scenario 2*, the weights for the NEM were calculated with an additional $A - C_1$ interaction term in the mediator model while the structural model was the same as in (16). For *Interaction Scenarios 3 and 4*, the NEM was implemented as in the main scenarios, except for an additional $A - A^*$ interaction term in the structural model (10). For the Exact approach, an $A - M$ interaction term was included in the outcome model for each interaction scenario investigated. Note that the software implementation provided in Samoilenko and Lefebvre (2021) and Caubet et al. (2022) for the Exact approach does not allow for exposure-covariate interaction terms in neither the outcome nor the mediator models. In the exact regression-based approach of Doretti et al. (2022), formulae for the natural effects that allow for exposure-covariate interactions are provided, but these are not, to our knowledge, implemented in software at this time.

3.4. Results - main scenarios

The results obtained for the main scenarios are presented in Tables 1 to 3, as well as in Table A.1 of Appendix A.3. The natural effects estimated using the proposed t-link mediation model were in agreement with the true values of the NDE and NIE specified for the three scenarios, with smaller biases and standard deviations observed with increased sample size n. The corresponding coverage probabilities of the 95% credible intervals for the NDE and NIE varied between 0.938 and 0.959, all scenarios and sample sizes combined. The NEM and Exact approaches returned NDE and NIE values that were also relatively close to the reference values. For the NDE, we found that the t-link approach was less biased and variable as compared to the other two approaches, at all sample sizes. We remarked that the NEM and Exact approaches behaved very similarly for the NDE in *Main Scenario 2*, while the Exact approach was found between the t-link and NEM approaches in *Main Scenarios 1 and 3* regarding bias and standard deviation values for this effect. Except in *Main Scenario 1* at the smallest sample size, we observed that the t-link and Exact approaches yielded similar results for the NIE. For this effect, the NEM approach was found generally more biased, but less variable than the other two approaches. As n increased, the NEM coverage probability for the NIE was observed to decrease since the variance decreased but the bias did not. For the total effect, the t-link and NEM approaches were found to behave similarly, especially at the largest sample size where the approaches appeared equivalent. With respect to credible/confidence intervals, the t-link approach generally yielded the smallest or close to smallest mean interval length as compared to the NEM and Exact approaches when the coverage probability for these last two approaches was near the nominal 0.95. This was especially noticeable for the smallest sample size in each scenario.

Table 1
Main Scenario 1. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 100$, $n = 250$ or $n = 1000$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 100$								
NDE_{t-link}	1.323	1.546	0.223	16.9	0.755	0.787	94.7 (3.246)	–
NDE_{NEM}	1.323	1.634	0.311	23.5	0.937	0.987	93.9 (4.987)	93.6 (3.368)
NDE_{Exact}	1.323	1.574	0.251	19.0	0.796	0.835	94.1 (4.355)	95.0 (3.052)
NIE_{t-link}	2.117	2.032	-0.085	-4.0	0.569	0.576	94.2 (2.516)	–
NIE_{NEM}	2.117	2.117	0.001	0.0	0.637	0.637	96.0 (3.169)	92.6 (2.417)
NIE_{Exact}	2.117	2.264	0.148	7.0	0.710	0.725	96.4 (3.794)	95.4 (2.839)
TE_{t-link}	2.800	3.138	0.338	12.1	1.615	1.650	95.3 (6.974)	–
TE_{NEM}	2.800	3.333	0.533	19.0	1.866	1.941	94.9 (11.37)	95.2 (7.422)
TE_{Exact}	2.800	3.477	0.677	24.2	1.872	1.990	94.3 (11.80)	95.4 (7.767)
$n = 250$								
NDE_{t-link}	1.323	1.413	0.090	6.8	0.383	0.393	95.2 (1.598)	–
NDE_{NEM}	1.323	1.488	0.165	12.5	0.431	0.461	94.2 (1.850)	94.2 (1.697)
NDE_{Exact}	1.323	1.465	0.142	10.8	0.392	0.417	94.4 (1.671)	95.0 (1.578)
NIE_{t-link}	2.117	2.081	-0.035	-1.7	0.374	0.376	94.4 (1.509)	–
NIE_{NEM}	2.117	2.021	-0.096	-4.5	0.341	0.354	93.3 (1.438)	90.6 (1.314)
NIE_{Exact}	2.117	2.141	0.024	1.2	0.384	0.385	95.2 (1.641)	94.4 (1.505)
TE_{t-link}	2.800	2.942	0.142	5.1	0.855	0.867	95.8 (3.596)	–
TE_{NEM}	2.800	2.972	0.172	6.2	0.879	0.896	95.3 (3.950)	95.9 (3.588)
TE_{Exact}	2.800	3.115	0.315	11.2	0.923	0.975	94.5 (4.133)	95.3 (3.806)
$n = 1000$								
NDE_{t-link}	1.323	1.343	0.021	1.6	0.178	0.179	94.7 (0.701)	–
NDE_{NEM}	1.323	1.420	0.097	7.3	0.193	0.216	93.0 (0.780)	93.2 (0.763)
NDE_{Exact}	1.323	1.407	0.085	6.4	0.182	0.200	93.5 (0.722)	94.0 (0.716)
NIE_{t-link}	2.117	2.105	-0.011	-0.5	0.190	0.191	94.0 (0.734)	–
NIE_{NEM}	2.117	1.987	-0.130	-6.1	0.163	0.208	86.1 (0.633)	83.8 (0.619)
NIE_{Exact}	2.117	2.091	-0.026	-1.2	0.181	0.183	93.5 (0.707)	93.3 (0.698)
TE_{t-link}	2.800	2.829	0.029	1.0	0.407	0.408	95.3 (1.624)	–
TE_{NEM}	2.800	2.813	0.013	0.5	0.398	0.398	95.7 (1.629)	95.2 (1.593)
TE_{Exact}	2.800	2.938	0.138	4.9	0.422	0.444	94.4 (1.718)	94.9 (1.688)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

3.5. Results - no mediation scenario

The results for the *No Mediation Scenario* are presented in Table 4 and Table A.2 of Appendix A.3. As seen in Table 4, relatively few differences in bias across the three approaches (t-link, NEM and Exact) were observed for NIE estimators on the OR scale under the no mediation scenario. For this effect, a slightly larger relative bias was nonetheless found for the t-link approach as compared to the NEM and Exact approaches at all sample sizes (relative bias values below 3% for the t-link). For the NDE, the t-link estimator was however noticeably less biased and variable than the other two approaches when $n = 100, 250$.

3.6. Results - interaction scenarios

Results for the t-link mediation approach as well as the NEM and Exact approaches for *Interaction Scenarios 1 to 3* are presented in Tables 5–7. For the sake of space, the results for *Interaction Scenario 4* are presented in Table A.3 of Appendix A.3. For *Interaction Scenarios 1 and 2*, additional results for the t-link approach when the baseline covariate C_1 is either fixed to 0 or 1 are presented in Tables A.4–A.5 of Appendix A.3.

For the t-link approach in *Interaction Scenarios 1 and 2*, we obtained relatively small biases and appropriate coverage probabilities for all sample sizes. These results are not unexpected given that the correct working models with interaction were fitted for these two scenarios. We also observed that the t-link approach offered increased numerical stability as compared to the NEM approach for the smallest sample size ($n = 250$) in these scenarios. For *Interaction Scenarios 3 and*

Table 2
Main Scenario 2. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	2.530	2.718	0.188	7.4	0.930	0.948	95.7 (3.819)	–
NDE_{NEM}	2.530	2.848	0.318	12.6	1.001	1.050	94.4 (4.654)	95.3 (4.000)
NDE_{Exact}	2.530	2.843	0.314	12.4	0.994	1.043	94.4 (4.549)	94.8 (3.977)
NIE_{t-link}	1.186	1.180	-0.007	-0.5	0.121	0.121	94.8 (0.495)	–
NIE_{NEM}	1.186	1.147	-0.039	-3.3	0.104	0.111	91.9 (0.444)	87.3 (0.395)
NIE_{Exact}	1.186	1.171	-0.015	-1.3	0.124	0.125	94.2 (0.532)	90.8 (0.466)
TE_{t-link}	3.000	3.208	0.208	6.9	1.070	1.090	95.2 (4.380)	–
TE_{NEM}	3.000	3.245	0.245	8.2	1.112	1.139	94.9 (5.168)	95.4 (4.450)
TE_{Exact}	3.000	3.303	0.303	10.1	1.128	1.168	94.3 (5.182)	95.3 (4.514)
$n = 500$								
NDE_{t-link}	2.530	2.607	0.078	3.1	0.604	0.609	95.3 (2.440)	–
NDE_{NEM}	2.530	2.720	0.190	7.5	0.639	0.666	94.1 (2.729)	95.0 (2.555)
NDE_{Exact}	2.530	2.721	0.191	7.6	0.637	0.665	94.3 (2.688)	94.8 (2.544)
NIE_{t-link}	1.186	1.182	-0.004	-0.3	0.083	0.083	95.3 (0.336)	–
NIE_{NEM}	1.186	1.145	-0.041	-3.5	0.069	0.080	90.2 (0.289)	86.4 (0.271)
NIE_{Exact}	1.186	1.164	-0.021	-1.8	0.081	0.083	94.1 (0.335)	90.9 (0.315)
TE_{t-link}	3.000	3.085	0.085	2.8	0.697	0.702	95.9 (2.808)	–
TE_{NEM}	3.000	3.104	0.104	3.5	0.711	0.719	95.6 (3.038)	95.8 (2.844)
TE_{Exact}	3.000	3.157	0.157	5.2	0.722	0.738	95.2 (3.051)	95.9 (2.885)
$n = 1500$								
NDE_{t-link}	2.530	2.543	0.013	0.5	0.323	0.324	95.8 (1.320)	–
NDE_{NEM}	2.530	2.650	0.120	4.7	0.342	0.363	93.8 (1.420)	94.2 (1.388)
NDE_{Exact}	2.530	2.653	0.123	4.9	0.341	0.362	93.7 (1.408)	94.1 (1.381)
NIE_{t-link}	1.186	1.186	0.000	0.0	0.047	0.047	94.9 (0.188)	–
NIE_{NEM}	1.186	1.144	-0.042	-3.6	0.038	0.057	82.2 (0.156)	78.6 (0.153)
NIE_{Exact}	1.186	1.162	-0.024	-2.0	0.044	0.051	91.1 (0.180)	89.7 (0.176)
TE_{t-link}	3.000	3.017	0.017	0.6	0.380	0.380	95.4 (1.525)	–
TE_{NEM}	3.000	3.028	0.028	0.9	0.386	0.387	95.1 (1.581)	95.4 (1.545)
TE_{Exact}	3.000	3.080	0.080	2.7	0.391	0.399	94.9 (1.597)	95.2 (1.568)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

4, we observed relative biases generally exceeding 10% (in absolute value) when fitting the t-link model with main effect terms only. Unlike in *Interaction Scenarios 1 and 2*, the magnitude of biases did not decrease when increasing sample size n in *Interaction Scenarios 3 and 4*. In these scenarios, the coverage probabilities were also significantly smaller than the nominal 0.95 value. Recall that *Interaction Scenarios 3 and 4* are such that the true outcome model was specified using a logistic regression model conditional on the mediator and also featured an interaction between exposure and mediator. The biases for the t-link NDE and NIE estimators were found larger for *Interaction Scenarios 3* compared to *Interaction Scenario 4*; this is not surprising given that the former exhibited larger level of $A - M$ interaction than the latter (*Interaction Scenario 3*: $\theta_3 = -\log(2.8)$; *Interaction Scenario 4*: $\theta_3 = -\log(1.68)$). For these two scenarios, significantly smaller biases were obtained for the t-link approach when interaction terms $A - C_1$ and $A - C_2$ were included in the fitted mean outcome model (results not shown), but this inclusion did not fully mitigate the biases seen in Table 7 and Table A.3. As compared to the t-link and NEM approaches, the Exact approach was globally the most performant when considering all interaction scenarios investigated together. Finally, given reported results for the NEM approach with weighting in these scenarios, we decided to explore whether better performance would be achieved when using a flexible NEM approach based on machine learning techniques (Steen et al., 2017). Mixed results were obtained (see Appendix A.3 for details and results in Table A.6).

4. Sensitivity analyses for mediator-outcome unmeasured confounding

We previously mentioned that any $M - Y$ confounder unadjusted for in the t-link mediation model ought to contribute to the actual value of ρ, the residual correlation between the latent mediator and outcome variables (see Fig. 2). Our pro-

Table 3
Main Scenario 3. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	1.375	1.473	0.098	7.1	0.420	0.431	95.6 (1.776)	–
NDE_{NEM}	1.375	1.682	0.307	22.3	0.517	0.601	91.3 (2.315)	92.4 (2.046)
NDE_{Exact}	1.375	1.594	0.218	15.9	0.427	0.479	93.7 (1.880)	95.0 (1.721)
NIE_{t-link}	2.036	2.015	-0.021	-1.0	0.379	0.380	94.2 (1.509)	–
NIE_{NEM}	2.036	1.802	-0.234	-11.5	0.276	0.362	87.0 (1.170)	80.1 (1.054)
NIE_{Exact}	2.036	2.081	0.046	2.2	0.400	0.403	94.3 (1.741)	94.2 (1.560)
TE_{t-link}	2.800	2.970	0.170	6.1	0.911	0.927	95.8 (3.781)	–
TE_{NEM}	2.800	3.007	0.207	7.4	0.964	0.986	95.2 (4.443)	95.7 (3.867)
TE_{Exact}	2.800	3.299	0.499	17.8	1.030	1.144	92.6 (4.689)	93.7 (4.131)
$n = 500$								
NDE_{t-link}	1.375	1.420	0.045	3.3	0.283	0.287	95.3 (1.141)	–
NDE_{NEM}	1.375	1.635	0.260	18.9	0.345	0.432	87.8 (1.422)	88.9 (1.347)
NDE_{Exact}	1.375	1.564	0.189	13.7	0.293	0.348	90.4 (1.190)	92.1 (1.150)
NIE_{t-link}	2.036	2.023	-0.013	-0.6	0.267	0.267	93.8 (1.037)	–
NIE_{NEM}	2.036	1.773	-0.263	-12.9	0.186	0.322	73.0 (0.746)	66.8 (0.708)
NIE_{Exact}	2.036	2.031	-0.005	-0.2	0.265	0.265	94.3 (1.068)	93.5 (1.028)
TE_{t-link}	2.800	2.876	0.076	2.7	0.626	0.630	95.2 (2.461)	–
TE_{NEM}	2.800	2.890	0.090	3.2	0.644	0.651	94.8 (2.661)	94.9 (2.502)
TE_{Exact}	2.800	3.170	0.370	13.2	0.689	0.782	91.1 (2.839)	92.2 (2.677)
$n = 1500$								
NDE_{t-link}	1.375	1.394	0.018	1.3	0.156	0.157	94.7 (0.621)	–
NDE_{NEM}	1.375	1.608	0.232	16.9	0.186	0.298	75.2 (0.755)	75.7 (0.741)
NDE_{Exact}	1.375	1.546	0.171	12.4	0.159	0.233	81.2 (0.643)	83.1 (0.638)
NIE_{t-link}	2.036	2.023	-0.013	-0.6	0.149	0.150	94.8 (0.584)	–
NIE_{NEM}	2.036	1.756	-0.279	-13.7	0.102	0.297	31.3 (0.403)	26.8 (0.396)
NIE_{Exact}	2.036	2.001	-0.035	-1.7	0.144	0.148	93.9 (0.566)	93.7 (0.567)
TE_{t-link}	2.800	2.821	0.021	0.7	0.341	0.341	94.9 (1.347)	–
TE_{NEM}	2.800	2.821	0.021	0.8	0.345	0.346	94.6 (1.392)	94.6 (1.363)
TE_{Exact}	2.800	3.091	0.291	10.4	0.372	0.472	87.3 (1.490)	88.6 (1.461)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

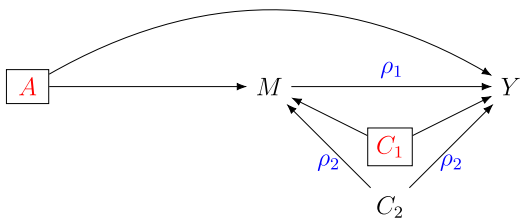

Fig. 2. Scenario in which covariates C_1 and C_2 are confounders the $M - Y$ relationship, but only C_1 is adjusted for in the model. Unadjustment for covariate C_2 contributes to the total residual correlation ρ through the open confounding fork $M \leftarrow C_2 \rightarrow Y$.

posed model is therefore expected to have a built-in capacity to evaluate the impact of violating the no-unmeasured $M - Y$ confounder assumption through parameter ρ. In the causal mediation analysis framework, sensitivity analysis procedures for unmeasured mediator-outcome confounding or more general unmeasured confounding (that is, either mediator-outcome, exposure-mediator or exposure-outcome confounding) based on a residual correlation parameter have previously been introduced (e.g., Imai et al. (2010b); Albert and Wang (2015); Lindmark et al. (2018)). As exemplified below, we propose a sensitivity analysis procedure for mediator-outcome confounding based on the residual correlation ρ.

We first illustrate the impact of omitting a common cause of M and Y in the t-link working model. More precisely, in two sensitivity analysis scenarios, we compared results obtained by correctly fitting a t-link model in which all $M - Y$

Table 4
No Mediation Scenario. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 100$, $n = 250$, or $n = 1000$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 100$								
NDE_{t-link}	2.800	3.232	0.432	15.4	1.770	1.822	94.6 (7.595)	–
NDE_{NEM}	2.800	3.520	0.720	25.7	2.148	2.266	93.5 (12.75)	93.6 (8.432)
NDE_{Exact}	2.800	3.441	0.641	22.9	2.015	2.114	93.7 (11.17)	94.3 (8.195)
NIE_{t-link}	1.000	1.026	0.026	2.6	0.169	0.171	96.3 (0.771)	–
NIE_{NEM}	1.000	1.008	0.008	0.8	0.191	0.192	94.2 (0.864)	96.1 (0.680)
NIE_{Exact}	1.000	1.012	0.012	1.2	0.196	0.196	95.5 (0.902)	97.4 (0.739)
TE_{t-link}	2.800	3.247	0.447	16.0	1.698	1.756	94.5 (7.231)	–
TE_{NEM}	2.800	3.422	0.622	22.2	1.955	2.051	94.3 (11.40)	94.4 (7.848)
TE_{Exact}	2.800	3.377	0.577	20.6	1.894	1.980	94.4 (10.45)	94.8 (7.625)
$n = 250$								
NDE_{t-link}	2.800	2.891	0.091	3.2	0.942	0.946	94.6 (3.746)	–
NDE_{NEM}	2.800	3.007	0.207	7.4	1.027	1.048	94.5 (4.373)	94.3 (3.948)
NDE_{Exact}	2.800	2.984	0.184	6.6	0.992	1.009	94.3 (4.171)	94.6 (3.871)
NIE_{t-link}	1.000	1.022	0.022	2.2	0.109	0.111	95.1 (0.456)	–
NIE_{NEM}	1.000	1.003	0.003	0.3	0.106	0.106	94.9 (0.440)	95.7 (0.403)
NIE_{Exact}	1.000	1.005	0.005	0.5	0.111	0.111	95.0 (0.463)	95.9 (0.430)
TE_{t-link}	2.800	2.930	0.130	4.7	0.915	0.924	94.5 (3.584)	–
TE_{NEM}	2.800	2.982	0.182	6.5	0.969	0.986	94.0 (4.046)	94.3 (3.693)
TE_{Exact}	2.800	2.968	0.168	6.0	0.949	0.964	93.7 (3.933)	94.6 (3.648)
$n = 1000$								
NDE_{t-link}	2.800	2.774	-0.026	-0.9	0.436	0.437	94.1 (1.686)	–
NDE_{NEM}	2.800	2.844	0.044	1.6	0.454	0.456	94.5 (1.792)	94.2 (1.751)
NDE_{Exact}	2.800	2.838	0.038	1.4	0.449	0.450	94.4 (1.753)	94.2 (1.726)
NIE_{t-link}	1.000	1.019	0.019	1.9	0.053	0.057	94.4 (0.218)	–
NIE_{NEM}	1.000	1.002	0.002	0.2	0.049	0.049	95.3 (0.198)	95.4 (0.193)
NIE_{Exact}	1.000	1.003	0.003	0.3	0.052	0.052	95.7 (0.209)	96.0 (0.205)
TE_{t-link}	2.800	2.821	0.021	0.8	0.419	0.420	94.2 (1.621)	–
TE_{NEM}	2.800	2.842	0.042	1.5	0.426	0.428	94.8 (1.687)	94.8 (1.650)
TE_{Exact}	2.800	2.839	0.039	1.4	0.426	0.428	94.4 (1.641)	94.6 (1.641)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

confounders were present to those obtained when omitting one of the confounders in the model. In these analyses, only samples of size $n = 250$ were considered.

The first sensitivity analysis scenario (*Scenario SA1*) considered a strong, possibly missing $M - Y$ confounder. In accordance with the DAG presented in Fig. 2, the regression equations and correlation value of the true generating data model were:

$$\mu^Y = -\log(2) + \log(2.8)A - \log(1.5)C_1 + \log(3)C_2,$$
$$\mu^M = -\log(2) + \log(4.5)A - \log(2)C_1 + \log(3)C_2,$$
$$\rho = 0.3,$$

where A and C_1 were Bernoulli random variables with $p = 0.5$, and C_2, the potentially missing confounding variable, was a standard normal random variable.

The second sensitivity analysis scenario (*Scenario SA2*) considered a moderately strong $M - Y$ confounder (C_2) whose omission induced an association in the opposite direction as compared to the residual dependency of the true generating model:

$$\mu^Y = -\log(2) + \log(2.8)A + \log(3)C_1 + \log(1.5)C_2,$$
$$\mu^M = -\log(2) + \log(4.5)A + \log(3)C_1 - \log(2)C_2,$$
$$\rho = 0.3,$$

Table 5
Interaction Scenario 1. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	2.682	2.959	0.277	10.3	1.045	1.081	94.8 (4.129)	–
NDE_{NEM}	2.682	3.101	0.419	15.6	1.153	1.227	93.5 (549.1)	94.8 (4.327)
NDE_{Exact}	2.682	2.847	0.165	6.2	0.953	0.967	94.7 (4.129)	94.8 (3.711)
NIE_{t-link}	1.501	1.485	-0.016	-1.1	0.203	0.204	93.9 (0.805)	–
NIE_{NEM}	1.501	1.442	-0.058	-3.9	0.180	0.189	91.6 (0.733)	87.5 (0.672)
NIE_{Exact}	1.501	1.491	-0.010	-0.7	0.203	0.203	94.1 (0.845)	92.1 (0.768)
TE_{t-link}	4.025	4.409	0.384	9.5	1.609	1.654	94.5 (6.277)	–
TE_{NEM}	4.025	4.448	0.423	10.5	1.686	1.739	93.8 (814.7)	94.6 (6.315)
TE_{Exact}	4.025	4.217	0.192	4.8	1.447	1.459	94.5 (6.300)	95.2 (5.616)
$n = 500$								
NDE_{t-link}	2.682	2.803	0.122	4.5	0.662	0.673	94.0 (2.548)	–
NDE_{NEM}	2.682	2.947	0.265	9.9	0.716	0.764	93.0 (2.976)	93.8 (2.717)
NDE_{Exact}	2.682	2.742	0.060	2.3	0.630	0.633	94.0 (2.515)	94.0 (2.404)
NIE_{t-link}	1.501	1.495	-0.006	-0.4	0.139	0.139	95.1 (0.557)	–
NIE_{NEM}	1.501	1.437	-0.064	-4.3	0.118	0.134	91.4 (0.484)	87.2 (0.462)
NIE_{Exact}	1.501	1.480	-0.020	-1.4	0.134	0.135	94.3 (0.546)	92.9 (0.523)
TE_{t-link}	4.025	4.203	0.178	4.4	1.037	1.052	93.7 (3.905)	–
TE_{NEM}	4.025	4.226	0.201	5.0	1.053	1.072	93.7 (4.343)	93.9 (3.949)
TE_{Exact}	4.025	4.051	0.026	0.6	0.964	0.965	93.7 (3.816)	93.4 (3.629)
$n = 1500$								
NDE_{t-link}	2.682	2.710	0.028	1.0	0.352	0.353	94.6 (1.354)	–
NDE_{NEM}	2.682	2.855	0.173	6.5	0.381	0.419	92.3 (1.496)	92.9 (1.457)
NDE_{Exact}	2.682	2.686	0.005	0.2	0.341	0.341	94.9 (1.337)	94.8 (1.317)
NIE_{t-link}	1.501	1.498	-0.003	-0.2	0.081	0.081	94.8 (0.314)	–
NIE_{NEM}	1.501	1.429	-0.072	-4.8	0.067	0.098	81.1 (0.264)	78.1 (0.260)
NIE_{Exact}	1.501	1.469	-0.032	-2.1	0.076	0.082	92.1 (0.295)	90.4 (0.292)
TE_{t-link}	4.025	4.060	0.035	0.9	0.542	0.543	94.4 (2.075)	–
TE_{NEM}	4.025	4.076	0.051	1.3	0.551	0.553	94.2 (2.164)	94.5 (2.103)
TE_{Exact}	4.025	3.942	-0.083	-2.1	0.511	0.518	94.5 (2.003)	94.1 (1.973)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

where variables A, C_1 and C_2 were generated the same way as in *Scenario SA1*.

For both *Scenario SA1* and *Scenario SA2*, we obtained NDE and NIE estimates on the OR scale as well as ρ estimates, where all estimates were based on the complete and reduced models. In these scenarios, the bias of the NDE and NIE estimators was relatively small when the two confounders were included in the model. Moreover, the bias in the estimation of the residual correlation ρ was also small. In each scenario, we obtained a large relative bias for ρ when considering the reduced model which omitted the confounder C_2. For example, in *Scenario SA1*, the relative bias for ρ was 66.6% (true $\rho = 0.30$; average estimated $\rho = 0.50$). In this scenario, the omission of C_2 yielded an underestimation of the NDE and an overestimation of the NIE.

We then reestimated the natural effects by adding increment Δ_ρ to the ρ sampled at every iteration of the MCMC algorithm obtained from the reduced models ($\Delta_\rho = -0.20$ in *Scenario SA1* and $\Delta_\rho = 0.14$ in *Scenario SA2*). Any such modified ρ value greater than 1 was truncated to 1, and parallely for a value less than -1. In each of these scenarios, Δ_ρ was selected to approximately equal the difference between the true ρ value and the average ρ estimate from the reduced model. To provide an additional understanding of the results obtained from these sensitivity analyses, estimated natural effects were also compared against the true effects that were computed by marginalizing over the C_2 covariate. This was done in order to compare NDE and NIE estimates to corresponding true effects that only conditioned on the observed covariates (here C_1).

All results for the first sensitivity analysis scenario (*Scenario SA1*) are found in Table 8. When adding the pertubation $\Delta_\rho = -0.20$ to the ρ value in the chain pertaining to the reduced model, the natural effects estimates were closer to the true values, as compared to the reduced model without post-processing. Coverage probabilities from the post-processed analysis were also closer to the nominal 0.95 value. Post-processed estimates were in fact particularly close to the true

Table 6
Interaction Scenario 2. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	2.762	2.969	0.207	7.5	0.915	0.938	94.9 (3.765)	–
NDE_{NEM}	2.762	3.045	0.282	10.2	0.988	1.028	94.9 (15.64)	95.2 (3.939)
NDE_{Exact}	2.762	3.037	0.275	9.9	0.953	0.991	94.6 (4.209)	94.9 (3.843)
NIE_{t-link}	1.086	1.081	-0.005	-0.5	0.067	0.067	94.2 (0.278)	–
NIE_{NEM}	1.086	1.086	0.000	0.0	0.078	0.078	96.5 (0.359)	94.2 (0.308)
NIE_{Exact}	1.086	1.085	-0.001	-0.1	0.067	0.067	94.3 (0.286)	88.9 (0.254)
TE_{t-link}	3.000	3.227	0.227	7.6	0.988	1.013	95.4 (4.074)	–
TE_{NEM}	3.000	3.295	0.295	9.8	1.058	1.098	94.8 (17.20)	95.2 (4.231)
TE_{Exact}	3.000	3.288	0.288	9.6	1.028	1.068	94.9 (4.555)	95.4 (4.157)
$n = 500$								
NDE_{t-link}	2.762	2.814	0.051	1.9	0.612	0.615	94.3 (2.397)	–
NDE_{NEM}	2.762	2.854	0.092	3.3	0.644	0.650	94.4 (2.618)	94.1 (2.473)
NDE_{Exact}	2.762	2.865	0.103	3.7	0.631	0.639	94.7 (2.540)	94.2 (2.442)
NIE_{t-link}	1.086	1.085	-0.002	-0.1	0.047	0.047	94.4 (0.184)	–
NIE_{NEM}	1.086	1.086	0.000	0.0	0.051	0.051	95.2 (0.220)	94.3 (0.203)
NIE_{Exact}	1.086	1.084	-0.002	-0.2	0.045	0.045	94.1 (0.181)	90.7 (0.171)
TE_{t-link}	3.000	3.062	0.062	2.1	0.667	0.670	94.6 (2.599)	–
TE_{NEM}	3.000	3.094	0.094	3.1	0.693	0.699	94.2 (2.809)	94.7 (2.658)
TE_{Exact}	3.000	3.103	0.103	3.4	0.685	0.692	94.4 (2.740)	94.7 (2.643)
$n = 1500$								
NDE_{t-link}	2.762	2.767	0.005	0.2	0.341	0.341	95.0 (1.322)	–
NDE_{NEM}	2.762	2.796	0.034	1.2	0.351	0.353	95.4 (1.383)	95.4 (1.358)
NDE_{Exact}	2.762	2.811	0.049	1.8	0.348	0.351	94.9 (1.363)	95.1 (1.346)
NIE_{t-link}	1.086	1.084	-0.002	-0.2	0.026	0.026	94.2 (0.100)	–
NIE_{NEM}	1.086	1.081	-0.005	-0.5	0.028	0.029	93.8 (0.113)	92.8 (0.109)
NIE_{Exact}	1.086	1.080	-0.006	-0.6	0.024	0.025	92.9 (0.095)	90.7 (0.094)
TE_{t-link}	3.000	3.003	0.003	0.1	0.368	0.368	95.0 (1.430)	–
TE_{NEM}	3.000	3.020	0.020	0.7	0.376	0.376	95.1 (1.479)	95.2 (1.453)
TE_{Exact}	3.000	3.034	0.034	1.1	0.375	0.376	95.1 (1.468)	95.3 (1.452)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval.

values obtained when marginalizing over the C_2 covariate. Similar results were obtained for *Scenario SA2* (see Table A.7 in Appendix A.3).

5. Application

It is known that survivors of childhood acute lymphoblastic leukemia (cALL) may experience severe long-term sequelae which are thought to be the result of treatment, disease, or both (Marcoux et al., 2017). These sequelae encompass a wide range of health complications including metabolic syndrome, cardiovascular problems, bone morbidity, and neuro-cognitive impairements. While the precise reasons for why survivors are predisposed to subsequent complications are still unclear, cranial radiation therapy and chemotherapy received are believed to be the source of several metabolic disturbances, which can lead to subsequent health conditions. This issue was explored in Caubet-Fernandez et al. (2019), which used the standard t-link model to assess the association between corticosteroid and cranial radiation exposures and four cardiometabolic complications (studied joint response variables were obesity, dyslipidemia, insulin resistance, and hypertension). In the current analysis, we wanted to shed light on the potential role of cranial radiation therapy (CRT) in the long-term risk of resistance to insulin while assuming obesity as a mediator. Indeed, while the etiology for the development for insulin resistance is still not fully understood in the general population, excess of visceral fat and obesity have long been identified as likely risk factors. Notably, the most recent report of the National Diabetes Statistics Report estimates, based on 2013-2016 data, that 89% of US adults diagnosed with diabetes are overweight or obese (Centers for Disease Control and Prevention, 2020).

Table 7
Interaction Scenario 3. Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the t-link mediation approach† (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	2.200	1.838	-0.363	-16.5	0.572	0.677	88.5 (2.343)	–
NDE_{NEM}	2.200	2.102	-0.098	-4.4	0.695	0.702	94.1 (3.040)	93.4 (2.780)
NDE_{Exact}	2.200	2.333	0.132	6.0	0.805	0.815	94.8 (3.631)	94.3 (3.263)
NIE_{t-link}	0.806	0.979	0.173	21.4	0.104	0.202	58.5 (0.440)	–
NIE_{NEM}	0.806	0.863	0.057	7.0	0.126	0.138	92.3 (0.502)	88.6 (0.465)
NIE_{Exact}	0.806	0.816	0.009	1.1	0.140	0.140	94.6 (0.565)	94.5 (0.536)
TE_{t-link}	1.774	1.774	0.000	0.0	0.506	0.506	94.8 (2.083)	–
TE_{NEM}	1.774	1.776	0.001	0.1	0.517	0.517	94.8 (2.230)	94.9 (2.098)
TE_{Exact}	1.774	1.845	0.075	4.2	0.551	0.556	95.3 (2.381)	95.5 (2.239)
$n = 500$								
NDE_{t-link}	2.200	1.826	-0.374	-17.0	0.407	0.553	82.2 (1.583)	–
NDE_{NEM}	2.200	2.083	-0.117	-5.3	0.488	0.502	93.7 (1.941)	92.9 (1.864)
NDE_{Exact}	2.200	2.308	0.107	4.9	0.568	0.578	94.2 (2.275)	94.8 (2.168)
NIE_{t-link}	0.806	0.973	0.167	20.7	0.078	0.184	37.2 (0.305)	–
NIE_{NEM}	0.806	0.854	0.047	5.9	0.085	0.097	90.1 (0.334)	87.6 (0.322)
NIE_{Exact}	0.806	0.805	-0.002	-0.2	0.097	0.097	93.6 (0.378)	94.2 (0.368)
TE_{t-link}	1.774	1.763	-0.011	-0.6	0.364	0.364	94.2 (1.414)	–
TE_{NEM}	1.774	1.760	-0.015	-0.8	0.365	0.365	94.4 (1.459)	94.6 (1.416)
TE_{Exact}	1.774	1.832	0.057	3.2	0.391	0.395	94.1 (1.549)	94.3 (1.506)
$n = 1500$								
NDE_{t-link}	2.200	1.779	-0.421	-19.2	0.219	0.475	56.6 (0.868)	–
NDE_{NEM}	2.200	2.016	-0.184	-8.4	0.256	0.316	88.4 (1.028)	87.9 (1.013)
NDE_{Exact}	2.200	2.219	0.019	0.9	0.294	0.294	95.4 (1.179)	95.4 (1.164)
NIE_{t-link}	0.806	0.972	0.166	20.5	0.043	0.171	2.9 (0.174)	–
NIE_{NEM}	0.806	0.856	0.049	6.1	0.048	0.069	81.9 (0.188)	78.6 (0.185)
NIE_{Exact}	0.806	0.809	0.002	0.3	0.053	0.053	95.3 (0.212)	95.2 (0.210)
TE_{t-link}	1.774	1.725	-0.050	-2.8	0.194	0.200	94.5 (0.780)	–
TE_{NEM}	1.774	1.720	-0.054	-3.1	0.195	0.203	94.2 (0.790)	94.3 (0.781)
TE_{Exact}	1.774	1.787	0.012	0.7	0.205	0.206	95.4 (0.834)	95.5 (0.827)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval. †: t-link model fitted with main effect terms only.

Our data come from the PETALE study, which was conducted at Sainte-Justine University Health Center (SJUHC) to characterize the most common late complications and identify associated predictive biomarkers in cALL survivors. Briefly, participants enrolled in the study ($n = 251$) were treated for cALL at SJUHC with the Dana Farber Cancer Institute protocols between January 1st 1987 and January 1st 2005. Survivors less than 19 years old at diagnosis, more than 5 years post diagnosis and free of relapse were invited to participate. At interview, which occurred between 2012 and 2016, participants underwent a complete clinical evaluation. We refer the reader to Marcoux et al. (2017) for a detailed presentation of the PETALE study and cohort.

Insulin resistance, as the binary outcome (Y), was defined by either high blood fasting glucose, high glycated hemoglobin, high homeostasis model assessment (HOMA-IR) [insulin (mIU/L) $\times$ glucose (mmol/L)/22.5] values or taking medication for diabetes. Obesity, as the binary mediator (M), was determined by presenting at least one of two factors at interview: obese according to body mass index (BMI) or high waist circumference. As our cohort mixes children and adult survivors, we point out the convenience of considering insulin resistance and obesity as binary variables since different cut-off values (pediatric and adult) were used to define them (see Caubet-Fernandez et al. (2019) for cut-off values for insulin resistance and obesity). Cranial radiation therapy was defined as a binary exposure variable (A) as most survivors who received CRT had an 18 Gray (Gy) cumulative dose, with none receiving over 20 Gy. A total of $n = 245$ survivors with complete information on outcome, mediator, treatment and selected adjustment covariates (*age at diagnosis*, *time since diagnosis*, *sex* and *white blood cell (WBC) count at diagnosis*) were used for the analyses. Descriptive statistics on this cohort of survivors are presented in Table B.8 of Appendix B.

Table 8
Sensitivity analysis (*Scenario SA1*). Natural direct and indirect effects (NDE, NIE, resp.) estimation on the odds ratio scale based on 2000 datasets of size $n = 250$ for complete and reduced t-link models, and reduced t-link model with pertubation $\Delta_\rho = -0.20$ at post-processing step.

Model	Estimand	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
Complete	NDE	2.122	2.278	0.157	7.4	0.748	0.764	94.8 (3.061)
	NIE	1.320	1.301	-0.018	-1.4	0.157	0.158	94.1 (0.632)
	ρ	0.300	0.283	-0.017	-5.6	0.108	0.110	95.0 (0.421)
Reduced	NDE	2.122	1.600	-0.522	-24.6	0.429	0.675	77.1 (1.750)
	NIE	1.320	1.480	0.160	12.2	0.177	0.235	85.9 (0.681)
	ρ	0.300	0.500	0.200	66.6	0.083	0.216	38.4 (0.325)
Reduced with Δ_ρ	NDE	2.122	1.888	-0.234	-11.0	0.503	0.555	91.1 (2.058)
	NIE	1.320	1.253	-0.067	-5.0	0.106	0.125	86.4 (0.419)
Reduced with $\Delta_\rho^\dagger$	NDE	1.830	1.888	0.058	3.2	0.503	0.506	95.6 (2.058)
	NIE	1.253	1.253	0.000	0.0	0.106	0.106	94.8 (0.419)

†: true values are conditional effects (at the mean of covariate C_1) marginalized over covariate C_2.
SD: standard deviation; RMSE: root mean squared error.
CP: coverage probability; CrI length : mean credible interval length.
ρ: residual correlation between latent outcome and mediator variables.

Table 9
Estimated regression coefficients for obesity (mediator) and insulin resistance (outcome)

	Obesity		Insulin resistance	
	Estimate†	95% CrI	Estimate†	95% CrI
Intercept	-1.410	(-2.489, -0.346)	-2.634	(-4.007, -1.324)
CRT	0.263	(-0.372, 0.909)	0.178	(-0.620, 0.993)
Age at diagnosis (years)	0.026	(-0.038, 0.090)	-0.017	(-0.100, 0.063)
Time since diagnosis (years)	0.039	(-0.019, 0.094)	0.063	(-0.004, 0.132)
Sex (male)	-0.854	(-1.420, -0.293)	-0.691	(-1.404, -0.008)
WBC count ($10^9/L$)	0.003	(-0.002, 0.008)	0.007	(0.002, 0.013)

†: posterior mean reported for point estimate; CrI: credible interval.

Table 10
Estimated conditional total effect (TE) and natural direct effect (NDE) of cranial radiation therapy on insulin resistance, with natural indirect effect (NIE) via obesity

	Females		Males	
	Estimate†	95% CrI	Estimate†	95% CrI
NDE^{OR}	1.043	(0.494, 2.146)	1.029	(0.485, 2.130)
NIE^{OR}	1.145	(0.823, 1.637)	1.163	(0.815, 1.683)
TE^{OR}	1.194	(0.538, 2.699)	1.194	(0.538, 2.699)
NDE^{RR}	1.035	(0.563, 1.891)	1.025	(0.524, 2.000)
NIE^{RR}	1.115	(0.856, 1.487)	1.143	(0.834, 1.597)
TE^{RR}	1.153	(0.607, 2.243)	1.171	(0.578, 2.445)
NDE^{RD}	0.005	(-0.107, 0.111)	0.001	(-0.074, 0.065)
NIE^{RD}	0.021	(-0.031, 0.077)	0.014	(-0.020,0.051)
TE^{RD}	0.026	(-0.099, 0.146)	0.015	(-0.067, 0.089)

†: posterior median reported for point estimate for the odds ratio (OR) and relative risk (RR) scales and posterior mean for point estimate for the risk difference (RD); CrI: credible interval.

The t-link model for mediation was fitted to the data to estimate the natural direct and indirect effects of CRT on insulin resistance on the OR, RR and RD scales. Adjustment covariates were included in the model as main effect terms and no interaction terms between treatment and covariates were considered. Bayesian estimation of regression and correlation coefficients were performed using the same prior distribution as presented in Section 3. We obtained 1 050 000 draws from the posterior distribution of $(\boldsymbol{\beta}, \mathbf{R})$ and discarded the first 50 000 as burn-in. After application of a thinning with factor of 100, 10 000 draws were available for inference. Posterior samples of conditional natural effects were obtained as described in the same section. For the continuous adjustment covariates *age at diagnosis*, *time since diagnosis*, and *WBC count at diagnosis*,

the natural effects were determined conditional on their empirical mean values. For the covariate *sex* its conditioning value was either fixed at 0 (female) or 1 (male); that is, we computed conditional effects according to female and male separately, and not on the average sex value. We emphasize that while two sets of conditional natural effects estimates were obtained, only one t-link model was fitted to the data.

Point estimates and 95% credible intervals (CrIs) for the regression coefficients of the t-link model are presented for obesity (mediator) and insulin resistance (outcome) in Table 9. While CRT was previously found positively associated with obesity among general childhood cancer survivors (e.g., Wilson et al. (2015)), there is insufficent evidence from our cohort of cALL survivors to conclude to an association between CRT and both obesity and insulin resistance in this specific population (log odds ratio for obesity associated with CRT: 0.263, 95% CrI: −0.372 to 0.909; log odds ratio for insulin resistance associated with CRT: 0.178, 95% CrI: −0.620 to 0.993). However, female sex was found positively associated with both obesity and insulin resistance, while WBC count at diagnosis was observed associated with insulin resistance. We estimated a large value for the correlation coefficient ($\rho = 0.699$, 95% CrI: 0.527 to 0.836); this indicates a high positive residual correlation (i.e., a correlation not explained by exposure and covariates) between obesity and insulin resistance. Under the assumption that no mediator-outcome confounders were excluded from the analysis, this high correlation suggests a strong causal link from obesity to insulin resistance. The point estimates and 95% CrIs for the conditional NDE and NIE are presented in Table 10, separately for females and males. Since we could not conclude for an association between CRT with neither outcome nor mediator, no mediation effects were found statistically significant. Nonetheless, for both females and males, the NDE was found close to the null while the NIE was only slightly smaller than the total effect (females: $NIE^{OR} = 1.145$, 95% CrI: 0.823 to 1.637; males: $NIE^{OR} = 1.163$, 95% CrI: 0.815 to 1.683). These results may suggest that the effect of CRT on insulin resistance is mostly mediated by obesity in cALL survivors.

Finally, we conducted a sensitivity analysis in order to evaluate the potential impact of an unmeasured $M-Y$ confounder on the natural effects estimates. To do so, we added a perturbation $\Delta\rho = \pm 0.2$ to each posterior draw for ρ and recomputed the NDE and NIE for each of the two values. For calibration, in our simulation presented in Section 4, a strong missing $M-Y$ confounder with association of $\beta = \log(3)$ with both the mediator and outcome yielded an increase of about 0.2 on the ρ estimate, on average. Results are presented in Tables B.9 and B.10 in Appendix B. As expected, the positive perturbation $\Delta\rho = 0.2$ increased the NIE estimates but decreased the NDE estimates, and vice-versa for the negative pertubation $\Delta\rho = -0.2$. In each case however, the natural effects estimates did not change drastically.

6. Discussion

Binary outcome variables are commonplace in many research areas, but are known to pose increased difficulty in mediation analyses. In this article, we introduced a new Bayesian causal mediation approach based on a joint model for a binary outcome and a binary mediator. Our novel approach utilizes the t-link model, a practical approximation of the multivariate logistic regression model introduced by O'Brien and Dunson (2004). Herein the t-link approach was successfully implemented on both simulated and real data with small or moderate sample sizes for the estimation of natural direct and indirect effects. In our application, we used the t-link model for mediation with a relatively common outcome variable, as about 17% of the PETALE cohort of cALL survivors were insulin resistant at interview. For binary outcomes, mediation analysis approaches can be classified according to whether or not an assumption regarding the rareness (or commonness) of the outcome is made. Conveniently, the t-link approach, as others such as the natural effect model (Lange et al., 2012; Steen et al., 2017) and exact regression-based (Samoilenko and Lefebvre, 2021) approaches, avoids making such assumptions.

As mentioned in the introduction, reasons for favoring a Bayesian framework in mediation are numerous, all of which incorporated or available in our approach. With the t-link approach to mediation, inference is conducted using MCMC methods which permits generating draws from the posterior distribution of the mediation effects (or any fonction thereof). As seen herein, mediation effects for a binary outcome were easily reported on all desired measure scales with proportion mediated, contrary to popular benchmark approaches from which results can only be readily reported on a single scale or requiring the specification of outcome models with different link functions to do so. For example, in the natural effect model approach of Lange et al. (2012) implemented in the R package `medflex` (Steen et al., 2017), the scale of the effects is tied to the link-function that is used for the outcome model. Moreover, the quasi-Bayesian approach of Imai et al. (2010a) implemented in the R package `mediation` (Tingley et al., 2014) returns natural effects on the risk difference scale only. To our knowledge, only the proposed t-link approach and the exact regression-based approach of Samoilenko and Lefebvre (2021) directly provide natural effects estimates on all scales (odds ratio, risk ratio, risk difference) on the basis of logistic models for the outcome and mediator. The t-link approach also allows the incorporation of relevant prior information on the associations between variables. While the prior distribution of the t-link model parameters was specified to be quite vague in our simulation and application, it is possible to use historical or other external data source to help determine an adequate informative prior distribution. Specifically, the t-link approach could first be fitted to historical data to calculate the posterior mean and variance of $(\boldsymbol{\beta}, \mathbf{R})$. These estimates could then be used directly as hyperparameter values for the prior $\pi(\boldsymbol{\beta}, \mathbf{R})$ or scaled using additional fixed or random hyperparameters to reflect the confidence one has on the historical data for bringing accurate information on the current data (Galharret and Philippe, 2019).

The t-link mediation approach permits sensitivity analyses to assess the impact of unmeasured confounding of the mediator-outcome relationship on the natural effects estimates. Our sensitivity analysis approach targets the residual correlation ρ between the latent mediator and outcome variables that underlies the joint binary model. The sensitivity analysis

is done as a post-processing step where the posterior distributions of the natural direct and indirect effects are estimated from the MCMC output in which the quantity Δ_ρ is added to the sampled values of ρ (the other parameters are unaltered). As in Imai et al. (2010b), we can then vary the value of the sensitivity parameter before recomputing the estimates. This provides a simple way to check for the impact of unmeasured confounding between the mediator and the outcome, and assess the robustness of empirical findings.

One constraint identified for the t-link mediation approach is that while the model allows for the absence of residual correlation between the outcome and mediator variables (that is $\rho = 0$), it does not allow for their (conditional) independance. This implies that the arrow $M \to Y$ should not be hypothesized missing in the context of the application. A strict recommendation would then be to avoid the t-link model if it is not reasonable to assume a causal link between the mediator and outcome. However, in our simulation, data featuring independent outcome and mediator yielded t-link indirect effect estimates very close to the null even with a very strong $A \to M$ arrow (coefficient of log(4.5)), suggesting that this negative attribute of the t-link model may have little impact in practice.

The proposed t-link approach allows the mean models for the outcome and mediator to be specified with exposure-covariate interactions so to capture mediated moderation or moderated mediation processes. However, it does not currently allow for the presence of an exposure-mediator interaction. This limitation should obviously be kept in mind in practice, so to restrict the use of the t-link model to contexts where no exposure-mediator interaction is expected or when a parcimonious model without such an interaction is desired. One way we could address this limitation in the future would be to extend the model to allow the residual correlation ρ to depend on the exposure variable A.

Our article has focused on the estimation of natural effects, where the statistical significance of direct, indirect and total effects was deduced from corresponding credible intervals. That is, an effect is said statistically significant if the reference value (1 for OR, RR and 0 for RD) is excluded from the credible interval. A more formal testing approach could be considered via the use of Bayes factors (Kass and Raftery, 1995; Nuijten et al., 2015). Acknowledging the composite nature of the null hypothesis (Barfield et al., 2017; Liu et al., 2021), testing for a null indirect effect could be accomplished by calculating the marginal likelihoods of four models: 1) a full t-link model; 2) a t-link model that excludes all regressors related to exposure in the mediator model; 3) a t-link model that forces the ρ coefficient to be zero; 4) a t-link model that excludes all regressors related to exposure in the mediator model and forces the ρ coefficient to be zero. Models 2), 3) and 4) encode (or most closely encode) the situations yielding no indirect effect, that is, either the exposure has no effect on the mediator, the mediator has no effect on the outcome, or both. Then the ratios of the marginal likelihood of the full model versus the marginal likelihoods of models 2, 3) or 4) could be obtained and assessed according known guidelines (Kass and Raftery, 1995). The marginal likelihood of the full model could also be compared to the sum of marginal likelihoods of models 2, 3) and 4) to account for total evidence for no mediation.

To conclude, given advantages and inconvenients discussed, we believe that our t-link approach for mediation is a worthwhile addition for practitioners conducting binary-binary mediation analyses.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; #2020-05473), Institute of Cancer Research (ICR) of the Canadian Institutes of Health Research (CIHR), in collaboration with C17 Council, Canadian Cancer Society (CCS), Cancer Research Society (CRS) and the Fonds de recherche Québec-Santé. Geneviève Lefebvre and Valérie Marcil are Research Scholars of the Fonds de recherche Québec-Santé. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Appendix A. Additional material on the simulation study

A.1. Data generation

We present a pseudocode describing the generation of one sample of size n according to the O'Brien and Dunson model (O'Brien and Dunson, 2004). The sample includes as variables the covariates C_1 and C_2, the exposure A, the mediator M and the outcome Y. The generative model features an interaction term between the exposure and covariate C_1 in the mean model for the mediator.

Pseudocode:

$n \leftarrow$ `Sample size`

$\boldsymbol{\beta}^Y \leftarrow (\beta_0^Y, \beta_A^Y, \beta_{C_1}^Y, \beta_{C_2}^Y)^T$

$\boldsymbol{\beta}^M \leftarrow (\beta_0^M, \beta_A^M, \beta_{AC_1}^M, \beta_{C_1}^M, \beta_{C_2}^M)^T$

$corr \leftarrow$ `Correlation value`

$C_1 \leftarrow$ `Random sample of size` n `from` $\mathrm{Ber}(p = 0.5)$

Table A.1
Main Scenario 1. Natural direct and indirect effects (NDE, NIE, resp.) estimation with the t-link mediation approach on the odds ratio (OR), risk ratio (RR) and risk difference (RD) scales, with corresponding proportion mediated (PM), using 2000 sample sizes of sizes $n = 100$, $n = 250$ and $n = 1000$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
$n = 100$							
NDE^{OR}	1.323	1.546	0.223	16.9	0.755	0.787	94.7 (3.246)
NIE^{OR}	2.117	2.032	-0.085	-4.0	0.569	0.576	94.2 (2.516)
PM^{OR}	0.728	0.673	-0.055	-7.6	0.310	0.315	96.1 (6.499)
NDE^{RR}	1.150	1.196	0.045	4.0	0.259	0.263	94.9 (1.059)
NIE^{RR}	1.326	1.286	-0.040	-3.0	0.159	0.164	93.6 (0.646)
PM^{RR}	0.668	0.620	-0.049	-7.3	0.369	0.372	96.0 (7.697)
NDE^{RD}	0.070	0.081	0.011	16.3	0.103	0.104	94.8 (0.399)
NIE^{RD}	0.174	0.151	-0.023	-13.0	0.058	0.062	93.2 (0.236)
PM^{RD}	0.714	1.734	1.020	142.9	35.18	35.20	96.0 (6.661)
$n = 250$							
NDE^{OR}	1.323	1.413	0.090	6.8	0.383	0.393	95.2 (1.598)
NIE^{OR}	2.117	2.081	-0.035	-1.7	0.374	0.376	94.4 (1.509)
PM^{OR}	0.728	0.736	0.008	1.1	0.226	0.227	96.6 (1.785)
NDE^{RR}	1.150	1.173	0.023	2.0	0.154	0.155	95.4 (0.621)
NIE^{RR}	1.326	1.308	-0.018	-1.3	0.103	0.105	94.3 (0.401)
PM^{RR}	0.668	0.683	0.015	2.3	0.272	0.273	96.4 (2.088)
NDE^{RD}	0.070	0.077	0.007	10.3	0.065	0.065	95.3 (0.254)
NIE^{RD}	0.174	0.164	-0.010	-5.5	0.040	0.041	93.8 (0.157)
PM^{RD}	0.714	0.816	0.102	14.3	1.747	1.750	96.5 (1.843)
$n = 1000$							
NDE^{OR}	1.323	1.343	0.021	1.6	0.178	0.179	94.7 (0.701)
NIE^{OR}	2.117	2.105	-0.011	-0.5	0.190	0.191	94.0 (0.734)
PM^{OR}	0.728	0.730	0.002	0.3	0.104	0.104	95.1 (0.423)
NDE^{RR}	1.150	1.156	0.005	0.5	0.075	0.076	95.0 (0.295)
NIE^{RR}	1.326	1.321	-0.005	-0.4	0.053	0.053	93.4 (0.199)
PM^{RR}	0.668	0.672	0.004	0.6	0.124	0.124	95.0 (0.502)
NDE^{RD}	0.070	0.072	0.002	2.4	0.032	0.032	94.8 (0.127)
NIE^{RD}	0.174	0.171	-0.003	-1.6	0.021	0.021	93.7 (0.080)
PM^{RD}	0.714	0.726	0.012	1.7	0.114	0.114	95.0 (0.446)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; CrI length: mean credible interval length.

$C_2 \leftarrow$ `Random sample of size` n `from` $\mathcal{N}(0,1)$

$A \leftarrow$ `Random sample of size` n `from` $\mathrm{Ber}(p_{\boldsymbol{c}})$, $p_{\boldsymbol{c}} = expit(\log(0.7) + \log(2)\,c_1 + \log(1.5)\,c_2)$ and $expit(\cdot) = exp(\cdot)/(1 + exp(\cdot))$

$TY \leftarrow$ `Combine by column` $(\mathbf{1}, A, C_1, C_2)$

$TM \leftarrow$ `Combine by column` $(\mathbf{1}, A, A * C_1, C_1, C_2)$

$\boldsymbol{\mu}_Y \leftarrow TY * \boldsymbol{\beta}^Y$

$\boldsymbol{\mu}_M \leftarrow TM * \boldsymbol{\beta}^M$

$\boldsymbol{\mu} \leftarrow$ `Combine by column` $(\boldsymbol{\mu}_Y, \boldsymbol{\mu}_M)$

$\mathbf{R} \leftarrow \begin{pmatrix} 1 & corr \\ corr & 1 \end{pmatrix}$

$\boldsymbol{e} \leftarrow$ `Random sample of size` n `from` $\mathcal{T}_{2,\nu}(\mathbf{0}, \boldsymbol{\Sigma} = \mathbf{R})$, $\nu = 7.3$

`Vectorized computations:`

$\boldsymbol{Z} \leftarrow \boldsymbol{\mu} + \log\left(\frac{\mathrm{F}(\boldsymbol{e})}{1-\mathrm{F}(\boldsymbol{e})}\right)$, (`F: CDF of univariate Student-`$t$ `with` ν `degrees of freedom`)

$YM \leftarrow \mathbb{1}(\boldsymbol{Z} > 0)$.

Table A.2
No Mediation Scenario. Natural direct and indirect effects (NDE, NIE, resp.) estimation with the t-link mediation approach on the odds ratio (OR), risk ratio (RR) and risk difference (RD) scales, with corresponding proportion mediated (PM), using 2000 samples of sizes $n = 100$, $n = 250$ or $n = 1000$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
$n = 100$							
NDE^{OR}	2.800	3.232	0.432	15.4	1.771	1.823	94.4 (7.592)
NIE^{OR}	1.000	1.026	0.026	2.6	0.169	0.171	96.2 (0.771)
PM^{OR}	0.000	0.014	0.014	–	0.158	0.158	97.8 (2.635)
NDE^{RR}	1.526	1.554	0.028	1.9	0.331	0.333	94.2 (1.340)
NIE^{RR}	1.000	1.007	0.007	0.7	0.048	0.048	96.2 (0.229)
PM^{RR}	0.000	0.016	0.016	–	0.117	0.118	97.8 (2.519)
NDE^{RD}	0.244	0.235	-0.009	-3.6	0.106	0.107	94.4 (0.411)
NIE^{RD}	0.000	0.005	0.005	–	0.033	0.034	96.2 (0.143)
PM^{RD}	0.000	0.008	0.008	–	3.264	3.264	97.8 (2.560)
ρ	0.000	-0.028	-0.028	–	0.173	0.175	95.3 (0.671)
$n = 250$							
NDE^{OR}	2.800	2.891	0.091	3.2	0.942	0.946	94.6 (3.746)
NIE^{OR}	1.000	1.022	0.022	2.2	0.109	0.111	95.1 (0.456)
PM^{OR}	0.000	0.016	0.016	–	0.115	0.116	95.5 (0.792)
NDE^{RR}	1.526	1.523	0.002	-0.1	0.202	0.202	94.2 (0.775)
NIE^{RR}	1.000	1.006	0.006	0.6	0.032	0.032	95.1 (0.135)
PM^{RR}	0.000	0.015	0.015	–	0.087	0.088	95.5 (0.692)
NDE^{RD}	0.244	0.235	-0.009	-3.7	0.070	0.070	94.3 (0.265)
NIE^{RD}	0.000	0.004	0.004	–	0.022	0.023	95.1 (0.090)
PM^{RD}	0.000	0.034	0.034	–	0.540	0.541	95.5 (0.748)
ρ	0.000	-0.021	-0.021	–	0.111	0.113	94.7 (0.441)
$n = 1000$							
NDE^{OR}	2.800	2.774	-0.026	-0.9	0.436	0.437	94.1 (1.686)
NIE^{OR}	1.000	1.019	0.019	1.9	0.053	0.057	94.4 (0.218)
PM^{OR}	0.000	0.017	0.017	–	0.053	0.056	94.4 (0.222)
NDE^{RR}	1.526	1.517	-0.009	-0.6	0.098	0.099	93.9 (0.375)
NIE^{RR}	1.000	1.006	0.006	0.6	0.016	0.017	94.4 (0.064)
PM^{RR}	0.000	0.013	0.013	–	0.038	0.041	94.4 (0.167)
NDE^{RD}	0.244	0.238	-0.006	-2.4	0.035	0.036	94.1 (0.134)
NIE^{RD}	0.000	0.004	0.004	–	0.011	0.012	94.4 (0.044)
PM^{RD}	0.000	0.017	0.017	–	0.048	0.051	94.4 (0.199)
ρ	0.000	-0.017	-0.017	–	0.057	0.059	94.2 (0.226)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; CrI length: mean credible interval length; ρ: residual correlation between latent outcome and mediator variables.

A.2. Computation of the natural effects

In this section, we present a pseudocode describing the algorithmic procedure for obtaining posterior samples of the conditional natural effects using the t-link model. We recall that, in this approach, the NDE and NIE are obtained by computing the nested counterfactuals $Y(a, M(a^*))$ using posterior draws of the model parameters, which we assume to be already available.

The procedure is encapsulated in the `Effects` function implemented in the R language. This function has six input parameters: $\boldsymbol{X}$, $\boldsymbol{beta}$, $\boldsymbol{corr}$, *delta_corr*, $\boldsymbol{int\_Y}$ and $\boldsymbol{int\_M}$. The parameter $\boldsymbol{X}$ is a two dimensional R list object. Its first component is a matrix whose ith row corresponds to the vector of outcome regressors associated with the ith observation, and the second component is the corresponding matrix for the mediator regressors. For example, for observation i, a vector of outcome regressors that includes interaction terms would be of the form: $\boldsymbol{X}_i^Y = (1, A_i, A_i\boldsymbol{C}_i, \boldsymbol{C}_i)$. The algorithm therefore allows to specify two different vectors of regressors, one for the outcome and the other for the mediator. The input parameter $\boldsymbol{beta}$ is a matrix whose kth row contains one posterior draw of the vector of regression coefficients; for example, the kth row of the $\boldsymbol{beta}$ matrix could be the vector $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_k^Y, \boldsymbol{\beta}_k^M) = (\beta_{0\,k}^Y, \beta_{A\,k}^Y, \boldsymbol{\beta}_{A\boldsymbol{C}\,k}^Y, \boldsymbol{\beta}_{\boldsymbol{C}\,k}^Y, \beta_{0\,k}^M, \beta_{A\,k}^M, \boldsymbol{\beta}_{A\boldsymbol{C}\,k}^M, \boldsymbol{\beta}_{\boldsymbol{C}\,k}^M)$. Input parameter $\boldsymbol{corr}$ is a column matrix containing the posterior draws of the correlation coefficient. The scalar input *delta_corr* allows the user to perform the sensitivity analysis for unmeasured confounding for the mediator-outcome relationship described herein by adding the perturbation *delta_corr* (Δ_ρ) to the posterior draws of the correlation. Finally, numeric vectors $\boldsymbol{int\_Y}$ and $\boldsymbol{int\_M}$ are two extra parameters specifying which covariates interact with the exposure in the mean models μ^Y and

Table A.3
Interaction Scenario 4. Natural direct and indirect effect (NDE, NIE, resp.) estimation with the t-link mediation approach† (t-link), natural effect model approach with weighting (NEM), and exact logistic regression-based approach (Exact) on the odds ratio scale. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (C(r)I length) perc/boot	CP (%) (C(r)I length) delta/robust
$n = 250$								
NDE_{t-link}	2.492	2.320	-0.172	-6.9	0.750	0.770	93.3 (3.028)	-
NDE_{NEM}	2.492	2.554	0.062	2.5	0.884	0.886	94.7 (3.821)	94.6 (3.437)
NDE_{Exact}	2.492	2.725	0.233	9.4	1.004	1.030	94.4 (4.429)	95.0 (3.917)
NIE_{t-link}	0.962	1.057	0.095	9.9	0.119	0.153	86.5 (0.491)	–
NIE_{NEM}	0.962	0.978	0.017	1.7	0.146	0.147	93.3 (0.584)	94.8 (0.528)
NIE_{Exact}	0.962	0.969	0.007	0.8	0.169	0.169	94.1 (0.684)	96.0 (0.633)
TE_{t-link}	2.397	2.422	0.025	1.1	0.718	0.719	95.4 (2.907)	–
TE_{NEM}	2.397	2.445	0.048	2.0	0.742	0.744	95.1 (3.186)	95.3 (2.952)
TE_{Exact}	2.397	2.561	0.164	6.8	0.789	0.805	94.3 (3.411)	95.1 (3.163)
$n = 500$								
NDE_{t-link}	2.492	2.253	-0.239	-9.6	0.504	0.558	90.4 (1.994)	–
NDE_{NEM}	2.492	2.464	-0.028	-1.1	0.573	0.574	94.4 (2.350)	94.2 (2.241)
NDE_{Exact}	2.492	2.622	0.130	5.2	0.641	0.654	94.2 (2.657)	94.6 (2.521)
NIE_{t-link}	0.962	1.056	0.094	9.8	0.084	0.126	77.7 (0.339)	–
NIE_{NEM}	0.962	0.973	0.011	1.1	0.099	0.099	94.4 (0.387)	94.9 (0.367)
NIE_{Exact}	0.962	0.961	-0.001	-0.1	0.113	0.113	95.1 (0.449)	96.0 (0.432)
TE_{t-link}	2.397	2.363	-0.034	-1.4	0.485	0.486	94.4 (1.930)	–
TE_{NEM}	2.397	2.373	-0.025	-1.0	0.493	0.494	94.5 (2.018)	94.2 (1.946)
TE_{Exact}	2.397	2.484	0.087	3.6	0.523	0.530	94.3 (2.152)	95.0 (2.080)
$n = 1500$								
NDE_{t-link}	2.492	2.196	-0.297	-11.9	0.272	0.402	80.7 (1.088)	–
NDE_{NEM}	2.492	2.389	-0.103	-4.1	0.304	0.320	93.6 (1.236)	93.1 (1.215)
NDE_{Exact}	2.492	2.529	0.037	1.5	0.339	0.341	95.4 (1.371)	95.6 (1.352)
NIE_{t-link}	0.962	1.056	0.095	9.8	0.049	0.106	47.9 (0.193)	–
NIE_{NEM}	0.962	0.973	0.011	1.1	0.053	0.054	95.2 (0.214)	95.4 (0.209)
NIE_{Exact}	0.962	0.962	0.000	0.0	0.062	0.062	94.6 (0.248)	95.4 (0.244)
TE_{t-link}	2.397	2.315	-0.082	-3.4	0.265	0.277	93.6 (1.063)	–
TE_{NEM}	2.397	2.317	-0.080	-3.3	0.268	0.279	93.7 (1.083)	93.7 (1.068)
TE_{Exact}	2.397	2.424	0.027	1.1	0.284	0.286	95.6 (1.150)	96.1 (1.139)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; C(r)I length: mean credible/confidence interval length; perc: equal-tailed credible interval; boot: percentile bootstrap interval. †: t-link model fitted with main effect terms only.

μ^M. Thus, if vector $\boldsymbol{int\_Y}$ is equal to $(2, 4)$, the user would be indicating that the second and fourth covariates in $\boldsymbol{C}$ interact with the exposure in the outcome model.

PSEUDOCODE: EFFECTS($\boldsymbol{X}$, $\boldsymbol{beta}$, $\boldsymbol{corr}$, $delta\_corr$, $\boldsymbol{int\_Y}$, $\boldsymbol{int\_M}$)

```
K ← Number of rows of beta matrix
IF type of corr = 'vector'
  THEN convert corr into a K × 1 matrix
ENDIF
IF delta_corr is not empty
  Add delta_corr to each entry of matrix corr
  Set entries of corr < −1 to −1
  Set entries of corr > 1 to 1
ENDIF
```

Table A.4
Interaction Scenario 1. Natural direct and indirect effects (NDE and NIE, resp.) and total effect (TE) estimation on the odds ratio scale with the t-link mediation approach when $C_1 = 0$ and $C_1 = 1$. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
$n = 250$							
$C_1 = 0$							
NDE	2.073	2.272	0.198	9.6	0.862	0.884	95.8 (3.714)
NIE	1.447	1.426	-0.021	-1.5	0.177	0.178	93.7 (0.705)
TE	3.000	3.264	0.264	8.8	1.291	1.318	95.5 (5.539)
$C_1 = 1$							
NDE	3.639	4.379	0.739	20.3	2.598	2.701	95.2 (10.62)
NIE	1.484	1.463	-0.021	-1.4	0.188	0.190	94.0 (0.748)
TE	5.400	6.434	1.034	19.2	3.906	4.040	94.9 (16.28)
$n = 500$							
$C_1 = 0$							
NDE	2.073	2.162	0.088	4.3	0.574	0.581	94.3 (2.343)
NIE	1.447	1.439	-0.008	-0.6	0.123	0.123	95.3 (0.491)
TE	3.000	3.126	0.126	4.2	0.879	0.888	94.1 (3.521)
$C_1 = 1$							
NDE	3.639	3.968	0.329	9.0	1.424	1.462	94.4 (5.721)
NIE	1.484	1.476	-0.008	-0.6	0.129	0.129	95.4 (0.518)
TE	5.400	5.870	0.470	8.7	2.162	2.212	94.2 (8.626)
$n = 1500$							
$C_1 = 0$							
NDE	2.073	2.112	0.039	1.9	0.322	0.324	95.4 (1.271)
NIE	1.447	1.443	-0.004	-0.3	0.072	0.072	94.6 (0.279)
TE	3.000	3.053	0.053	1.8	0.482	0.485	95.0 (1.908)
$C_1 = 1$							
NDE	3.639	3.695	0.056	1.5	0.728	0.730	94.0 (2.808)
NIE	1.484	1.480	-0.004	-0.3	0.075	0.076	94.5 (0.293)
TE	5.400	5.471	0.071	1.3	1.104	1.106	93.9 (4.223)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; CrI length: mean credible interval length.

```
FOR j ∈ {Y, M}:
  Set vector x_m^j of list x_m to the column mean values of the jth matrix of the list X.
  (Remark: Subscript m is used for 'mean' and not 'mediator')
  Set vector xh_m^j of list xh_m to the value of x_m^j
  Set vector xl_m^j of list xl_m to the value of x_m^j
  Set the exposure component of vector xh_m^j to 1
  Set the exposure component of vector xl_m^j to 0
ENDFOR
l_Y ← length of vector int_Y
l_M ← length of vector int_M
IF l_Y ≠ 0
  FOR i = 1, ..., l_Y
```

Table A.5
Interaction Scenario 2. Natural direct and indirect effects (NDE and NIE, resp.) estimation on the odds ratio scale with the t-link mediation approach when $C_1 = 0$ and $C_1 = 1$. Results based on 2000 datasets of size $n = 250$, $n = 500$ or $n = 1500$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
$n = 250$							
$C_1 = 0$							
NDE	2.923	3.125	0.203	6.9	0.960	0.981	94.9 (3.962)
NIE	1.026	1.029	0.003	0.3	0.055	0.055	97.9 (0.271)
TE	3.000	3.227	0.227	7.6	0.988	1.013	95.4 (4.074)
$C_1 = 1$							
NDE	2.603	2.823	0.220	8.4	0.891	0.918	95.0 (3.682)
NIE	1.152	1.140	-0.012	-1.1	0.108	0.109	94.4 (0.443)
TE	3.000	3.227	0.227	7.6	0.988	1.013	95.4 (4.074)
$n = 500$							
$C_1 = 0$							
NDE	2.923	2.974	0.051	1.8	0.649	0.651	94.6 (2.531)
NIE	1.026	1.027	0.001	0.1	0.039	0.039	95.7 (0.173)
TE	3.000	3.062	0.062	2.1	0.667	0.670	94.6 (2.599)
$C_1 = 1$							
NDE	2.603	2.660	0.057	2.2	0.592	0.594	94.4 (2.331)
NIE	1.152	1.149	-0.004	-0.3	0.076	0.077	94.2 (0.301)
TE	3.000	3.062	0.062	2.1	0.667	0.670	94.6 (2.599)
$n = 1500$							
$C_1 = 0$							
NDE	2.923	2.926	0.003	0.1	0.359	0.359	94.8 (1.395)
NIE	1.026	1.025	-0.001	-0.1	0.022	0.022	94.8 (0.090)
TE	3.000	3.003	0.003	0.1	0.368	0.368	95.0 (1.430)
$C_1 = 1$							
NDE	2.603	2.611	0.008	0.3	0.330	0.330	95.0 (1.282)
NIE	1.152	1.149	-0.003	-0.3	0.043	0.043	94.6 (0.167)
TE	3.000	3.003	0.003	0.1	0.368	0.368	95.0 (1.430)

SD: standard deviation; RMSE: root mean squared error; CP: coverage probability; CrI length: mean credible interval length.

Set the ith interaction component of $\boldsymbol{xh}_m^Y$ to the value of the component in $\boldsymbol{xh}_m^Y$ that corresponds to the covariate specified in the ith component of $\boldsymbol{int\_Y}$
Set the ith interaction component of $\boldsymbol{xl}_m^Y$ to 0

ENDFOR

ENDIF

IF $l\_M \neq 0$
FOR $i = 1, \ldots, l\_M$

Set the ith interaction component of $\boldsymbol{xh}_m^M$ to the value of the component in $\boldsymbol{xh}_m^M$ that corresponds to the covariate specified in the ith component of $\boldsymbol{int\_M}$
Set the ith interaction component of $\boldsymbol{xl}_m^M$ to 0

ENDFOR

ENDIF

Table A.6
Natural direct and indirect effects (NDE, NIE, resp.) and total effect (TE) estimation with the natural effect model approach with imputation based on an ensemble learner for *Interaction Scenarios 1 to 4*. Results based on 2000 datasets of size $n = 250$.

Effects	True value	Mean	Bias	Relative bias (%)	SD	RMSE
			Interaction Scenario 1			
NDE	2.682	2.847	0.165	6.15	1.046	1.059
NIE	1.501	1.428	-0.072	-4.81	0.177	0.191
TE	4.025	4.036	0.011	0.28	1.484	1.484
			Interaction Scenario 2			
NDE	2.762	2.986	0.224	8.1	0.968	0.994
NIE	1.086	1.073	-0.013	-1.2	0.069	0.070
TE	3.000	3.194	0.194	6.5	1.019	1.038
			Interaction Scenario 3			
NDE	2.200	1.816	-0.385	-17.5	0.678	0.773
NIE	0.806	0.926	0.119	14.8	0.110	0.163
TE	1.774	1.649	-0.125	-7.1	0.552	0.566
			Interaction Scenario 4			
NDE	2.492	2.340	-0.152	-6.1	0.825	0.839
NIE	0.962	1.009	0.047	4.9	0.099	0.110
TE	2.397	2.335	-0.062	-2.6	0.770	0.772

SD: standard deviation; RMSE: root mean squared error.

Table A.7
Sensitivity analysis (*Scenario SA2*). Natural direct and indirect effects (NDE, NIE, resp.) estimation on the odds ratio scale based on 2000 datasets of size $n = 250$ for complete and reduced models, and reduced model with pertubation $\Delta_\rho = 0.14$ at post-processing step.

Model	Estimand	True value	Mean	Bias	Relative bias (%)	SD	RMSE	CP (%) (CrI length)
Complete	NDE	2.038	2.171	0.133	6.5	0.668	0.681	95.1 (2.686)
	NIE	1.374	1.362	-0.012	-0.9	0.177	0.178	93.8 (0.705)
	ρ	0.300	0.288	-0.012	-3.8	0.109	0.109	95.1 (0.419)
Reduced	NDE	2.038	2.385	0.347	17.0	0.709	0.789	93.2 (2.854)
	NIE	1.374	1.182	-0.191	-13.9	0.122	0.227	67.3 (0.491)
	ρ	0.300	0.163	-0.137	-45.7	0.107	0.174	75.7 (0.412)
Reduced with Δ_ρ	NDE	2.038	2.111	0.073	3.6	0.631	0.636	95.5 (2.540)
	NIE	1.374	1.335	-0.038	-2.8	0.160	0.165	92.8 (0.638)
Reduced with $\Delta_\rho^\dagger$	NDE	2.050	2.111	0.061	3.0	0.631	0.634	95.5 (2.540)
	NIE	1.315	1.335	0.020	1.5	0.160	0.162	94.6 (0.638)

†: true values are conditional effects (at the mean of covariate C_1) marginalized over covariate C_2.
SD: standard deviation; RMSE: root mean squared error;
CP: coverage probability; CrI length : mean credible interval length;
ρ: residual correlation between latent outcome and mediator variables.

$cvt \leftarrow$ `length of vector` $\boldsymbol{x}_m^Y$ + `length of vector` $\boldsymbol{x}_m^M$

$M_m \leftarrow \begin{pmatrix} \boldsymbol{xh}_m^Y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{xh}_m^M \end{pmatrix}$

$E \leftarrow \mathbf{0}_{K,3}$

`FOR` $k = 1, \ldots, K$:

`Compute the scale matrix:`

$$\Sigma \leftarrow \delta^2 \begin{pmatrix} 1 & corr_k \\ corr_k & 1 \end{pmatrix} \quad \texttt{where} \quad \delta = \pi^2 \frac{(\nu-2)}{(3\nu)} \quad \texttt{with} \quad \nu = 7.3$$

Compute the location parameter:

$$\boldsymbol{\mu} \leftarrow M_m \begin{pmatrix} \boldsymbol{\beta}_k^Y \\ \boldsymbol{\beta}_k^M \end{pmatrix}$$

Compute probability $\mathrm{P}_k(M=1|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}})$:

$$p_1 \leftarrow \frac{\exp(\boldsymbol{x}\boldsymbol{h}_m^M \boldsymbol{\beta}_k^M)}{1+\exp(\boldsymbol{x}\boldsymbol{h}_m^M \boldsymbol{\beta}_k^M)}$$

Compute probability $\mathrm{P}_k(M=1|A=0, \boldsymbol{C}=\bar{\boldsymbol{c}})$:

$$p_0 \leftarrow \frac{\exp(\boldsymbol{x}\boldsymbol{l}_m^M \boldsymbol{\beta}_k^M)}{1+\exp(\boldsymbol{x}\boldsymbol{l}_m^M \boldsymbol{\beta}_k^M)}$$

Compute counterfactual probability

$$\begin{aligned}\mathrm{P}_k(Y(1, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}}) = \mathrm{P}_k(Y=1, M=1|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}}) \frac{\mathrm{P}_k(M=1|A=0, \boldsymbol{C}=\bar{\boldsymbol{c}})}{\mathrm{P}_k(M=1|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}})} \\ + \mathrm{P}_k(Y=1, M=0|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}}) \frac{\mathrm{P}_k(M=0|A=0, \boldsymbol{C}=\bar{\boldsymbol{c}})}{\mathrm{P}_k(M=0|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}})}:\end{aligned}$$

$$E_{k,1} \leftarrow \int_0^{+\infty}\int_0^{+\infty} \mathrm{St}(\boldsymbol{t}; \boldsymbol{\mu}, \Sigma, \tilde{\nu})\, \mathrm{d}\boldsymbol{t} \times \frac{p_0}{p_1} + \int_0^{+\infty}\int_{-\infty}^{0} \mathrm{St}(\boldsymbol{t}; \boldsymbol{\mu}, \Sigma, \tilde{\nu})\, \mathrm{d}\boldsymbol{t} \times \frac{1-p_0}{1-p_1},$$

where $\tilde{\nu}=7$ and $\boldsymbol{t}=(t_1, t_2)$

ENDFOR

Vectorized computation (without 'for' loop) of the counterfactual probabilities $\mathrm{P}_k(Y(0, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}}) = \mathrm{P}_k(Y=1|A=0, \boldsymbol{C}=\bar{\boldsymbol{c}}),\ k=1,\ldots,K$:

$$E_{k,2} \leftarrow \frac{\exp(\boldsymbol{x}\boldsymbol{l}_m^Y \boldsymbol{\beta}_k^Y)}{1+\exp(\boldsymbol{x}\boldsymbol{l}_m^Y \boldsymbol{\beta}_k^Y)}$$

Vectorized computation of counterfactual probabilities $\mathrm{P}_k(Y(1, M(1))=1|\boldsymbol{C}=\bar{\boldsymbol{c}}) = \mathrm{P}_k(Y=1|A=1, \boldsymbol{C}=\bar{\boldsymbol{c}}),\ k=1,\ldots,K$:

$$E_{k,3} \leftarrow \frac{\exp(\boldsymbol{x}\boldsymbol{h}_m^Y \boldsymbol{\beta}_k^Y)}{1+\exp(\boldsymbol{x}\boldsymbol{h}_m^Y \boldsymbol{\beta}_k^Y)}$$

$OR \leftarrow \mathbf{0}_{K,3}$
$RR \leftarrow \mathbf{0}_{K,3}$
$RD \leftarrow \mathbf{0}_{K,3}$

Vectorized computation of natural direct effects on odds ratio scale,

$$OR_{1,0|\bar{\boldsymbol{c}}_k}^{NDE} = \frac{\mathrm{P}_k(Y(1, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}})/(1-\mathrm{P}_k(Y(1, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}}))}{\mathrm{P}_k(Y(0, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}})/(1-\mathrm{P}_k(Y(0, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}}))}:$$

$$OR_{k,1} \leftarrow \frac{E_{k,1}/(1-E_{k,1})}{E_{k,2}/(1-E_{k,2})},\ k=1,\ldots,K$$

Vectorized computation of natural direct effects on risk ratio scale,

$$RR_{1,0|\bar{\boldsymbol{c}}_k}^{NDE} = \frac{\mathrm{P}_k(Y(1, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}})}{\mathrm{P}_k(Y(0, M(0))=1|\boldsymbol{C}=\bar{\boldsymbol{c}})}:$$

$$RR_{k,1} \leftarrow \frac{E_{k,1}}{E_{k,2}},\ k = 1, \ldots, K$$

Vectorized computation of direct effects on risk difference scale, $RD^{NDE}_{1,0|\bar{\boldsymbol{c}}_k} = \mathrm{P}_k(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}}) - \mathrm{P}_k(Y(0, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})$:

$$RD_{k,1} \leftarrow E_{k,1} - E_{k,2},\ k = 1, \ldots, K$$

Vectorized computation of natural indirect effects on odds ratio scale,

$$OR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k} = \frac{\mathrm{P}_k(Y(1, M(1)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})/(1 - \mathrm{P}_k(Y(1, M(1)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}}))}{\mathrm{P}_k(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})/(1 - \mathrm{P}_k(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}}))}:$$

$$OR_{k,2} \leftarrow \frac{E_{k,3}/(1 - E_{k,3})}{E_{k,1}/(1 - E_{k,1})},\ k = 1, \ldots, K$$

Vectorized computation of natural indirect effects on risk ratio scale,

$$RR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k} = \frac{\mathrm{P}_k(Y(1, M(1)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})}{\mathrm{P}_k(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})}:$$

$$RR_{k,2} \leftarrow \frac{E_{k,3}}{E_{k,1}},\ k = 1, \ldots, K$$

Vectorized computation of indirect effects on risk difference scale,

$$RD^{NIE}_{1,0|\bar{\boldsymbol{c}}_k} = \mathrm{P}_k(Y(1, M(1)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}}) - \mathrm{P}_k(Y(1, M(0)) = 1|\boldsymbol{C} = \bar{\boldsymbol{c}})$$

$$RD_{k,2} \leftarrow E_{k,3} - E_{k,1},\ k = 1, \ldots, K$$

Vectorized computation of proportions mediated on odds ratio scale,

$$OR^{PM}_{1,0|\bar{\boldsymbol{c}}_k} = \frac{\log(OR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k})}{\log(OR^{NDE}_{1,0|\bar{\boldsymbol{c}}_k}) + \log(OR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k})}:$$

$$OR_{k,3} \leftarrow \frac{\log(OR_{k,2})}{\log(OR_{k,1}) + \log(OR_{k,2})},\ k = 1, \ldots, K$$

Vectorized computation of proportions mediated on risk ratio scale,

$$RR^{PM}_{1,0|\bar{\boldsymbol{c}}_k} = \frac{\log(RR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k})}{\log(RR^{NDE}_{1,0|\bar{\boldsymbol{c}}_k}) + \log(RR^{NIE}_{1,0|\bar{\boldsymbol{c}}_k})}:$$

$$RR_{k,3} \leftarrow \frac{\log(RR_{k,2})}{\log(RR_{k,1}) + \log(RR_{k,2})},\ k = 1, \ldots, K$$

Vectorized computation of proportions mediated on risk difference scale,

$$RD^{PM}_{1,0|\bar{\boldsymbol{c}}_k} = \frac{RD^{NIE}_{1,0|\bar{\boldsymbol{c}}_k}}{RD^{NDE}_{1,0|\bar{\boldsymbol{c}}_k} + RD^{NIE}_{1,0|\bar{\boldsymbol{c}}_k}}:$$

$$RD_{k,3} \leftarrow \frac{RD_{k,2}}{RD_{k,1} + RD_{k,2}},\ k = 1, \ldots, K$$

A.3. Additional simulation results

This section of the Appendix presents additional results for the simulations described in Sections 3 and 4 of the manuscript.

The *Main Scenario 1* results for the NDE and NIE on all three scales (OR, RR and RD) with corresponding proportion mediated are presented for the t-link approach in Table A.1. As expected, the results were in agreement with the reference values. A very large variance for the proportion mediated on the RD scale was however obtained when $n = 100$, due to extreme values obtained for some datasets.

Additional results for the t-link mediation approach in the *No Mediation Scenario* are presented in Table A.2. The NDE and NIE were estimated with small or negligible biases when expressed on the RR or RD scale. For the OR scale, the relative bias of 15.4% observed for the NDE when $n = 100$ decreased to 3.2% when $n = 250$ and to less than 1% (in absolute value) when $n = 1000$. For this scale and all three sample sizes, the average NIE estimates returned by the t-link approach were close to the null value ($NIE^{OR} = 1$). The conditional independence between the outcome Y and mediator M led to average correlation coefficient ρ estimates near zero, with coverage probabilities of credible intervals enclosing the true ρ value ($\rho = 0$) close to the nominal level of 95%.

Results for the *Interaction Scenario 4* are found in Table A.3; these results are described along those for *Interaction Scenario 3* in Section 3.6. Results for the t-link mediation approach in the *Interaction Scenarios 1 and 2* at the two values of the binary covariate C_1 are presented in Tables A.4 and A.5, respectively. For *Interaction Scenario 1*, we observed biases that were relatively large for the smallest sample size ($n = 250$), especially when $C_1 = 1$, but the relative biases were below 2% at the largest sample size ($n = 1500$). For this scenario, we found that the value of the covariate C_1 influenced considerably the magnitude of the NDE, while the NIE hardly changed. This behavior corresponds to a mediated moderation scenario where there is only little indirect effect modification by covariate C_1, which is not unexpected given the absence of an interaction term $A - C_1$ in the mean model for the mediator in this scenario. In *Interaction Scenario 2*, the relative biases were smaller (in absolute values) than those obtained in *Interaction Scenario 1*. In this scenario, the value of covariate C_1 influenced the magnitude of the direct and indirect effect in opposite ways, which is line with a moderated mediation scenario (total effect unchanged, with true $TE = 3$ for both $C_1 = 0$ and $C_1 = 1$). For all interaction scenarios, we performed additional analyses for the NEM approach when $n = 250$. More precisely, we implemented the NEM with imputation using an ensemble learner, as implemented in the R package `SuperLearner` (Polley et al., 2021), to fit the imputation model (Steen et al., 2017). The results are found in Table A.6. As compared to the t-link approach, this alternative NEM approach yielded smaller bias for the NDE and TE in *Interaction Scenario 1*, but smaller bias was only obtained for the TE in *Interaction Scenario 2*. In *Interaction Scenarios 3 and 4*, the above NEM approach with imputation was markedly more biased than the NEM approach with weighting considered for these scenarios. In *Interaction Scenario 3*, the NEM approach with imputation yielded NDE and TE estimators that were more biased than those obtained from the t-link approach.

Results for the second sensitivity analysis scenario (*Scenario SA2*) are found in Table A.7; these results are described along those for *Scenario SA1* in Section 4.

Appendix B. Additional application results

Descriptive statistics on the PETALE cohort of survivors are presented in Table B.8.

Results for the sensitivity analysis to strong unmeasured confounding between obesity and insulin resistance in the real data example are presented in Tables B.9 and B.10.

Table B.8
Characteristics of the PETALE childhood acute lymphoblastic leukemia survivors cohort

	Full cohort ($n = 245$)	No CRT ($n = 100$)	CRT ($n = 145$)
Insulin resistance	42 (17.1%)	13 (13.0%)	29 (20.0%)
Obesity	79 (32.2%)	27 (27.0%)	52 (35.9%)
Sex (male)	121 (49.4%)	42 (42.0%)	79 (54.5%)
Age at diagnosis (years)	6.61 (4.60)	5.30 (3.41)	7.52 (5.09)
Time since diagnosis (years)	15.5 (5.16)	14.0 (4.73)	16.5 (5.23)
WBC count at diagnosis ($10^9/L$)	30.4 (53.2)	8.25 (7.80)	45.7 (64.6)

Results are presented as frequency and % (in parenthesis) for the binary variables and as mean and standard deviation (in parenthesis) for the continuous variables. CRT: cranial radiation therapy, WBC: white blood cell.

Table B.9
Sensitivity analysis with $\Delta\rho = 0.2$. Estimated conditional total effect (TE) and natural direct effect (NDE) of cranial radiation therapy on insulin resistance, with natural indirect effect (NIE) via obesity

Effects	Females		Males	
	Estimate†	95% CrI	Estimate†	95% CrI
NDE^{OR}	0.995	(0.470, 2.059)	0.966	(0.450, 2.026)
NIE^{OR}	1.196	(0.774, 1.883)	1.230	(0.756, 2.061)
TE^{OR}	1.194	(0.538, 2.699)	1.194	(0.538, 2.699)
NDE^{RR}	0.996	(0.538, 1.829)	0.969	(0.489, 1.910)
NIE^{RR}	1.155	(0.816, 1.683)	1.203	(0.781, 1.927)
TE^{RR}	1.153	(0.607, 2.243)	1.171	(0.578, 2.445)
NDE^{RD}	-0.001	(-0.117, 0.106)	-0.004	(-0.079, 0.063)
NIE^{RD}	0.027	(-0.041, 0.099)	0.019	(-0.029, 0.068)
TE^{RD}	0.026	(-0.099, 0.146)	0.015	(-0.067, 0.089)

†: posterior median reported for estimate for the odds ratio (OR) and relative risk (RR) scales and posterior mean for estimate for the risk difference (RD); CrI: credible interval.

Table B.10
Sensitivity analysis with $\Delta\rho = -0.2$. Estimated conditional total effect (TE) and natural direct effect (NDE) of cranial radiation therapy on insulin resistance, with natural indirect effect (NIE) via obesity

Effects	Females		Males	
	Estimate†	95% CrI	Estimate†	95% CrI
NDE^{OR}	1.090	(0.511, 2.272)	1.081	(0.507, 2.244)
NIE^{OR}	1.098	(0.873, 1.399)	1.108	(0.871, 1.410)
TE^{OR}	1.194	(0.538, 2.699)	1.194	(0.538, 2.699)
NDE^{RR}	1.072	(0.583, 1.978)	1.072	(0.547, 2.100)
NIE^{RR}	1.077	(0.896, 1.308)	1.095	(0.883, 1.359)
TE^{RR}	1.153	(0.607, 2.243)	1.171	(0.578, 2.445)
NDE^{RD}	0.011	(-0.104, 0.118)	0.005	(-0.071, 0.070)
NIE^{RD}	0.015	(-0.021, 0.055)	0.010	(-0.014, 0.035)
TE^{RD}	0.026	(-0.099, 0.146)	0.015	(-0.067, 0.089)

†: posterior median reported for estimate for the odds ratio (OR) and relative risk (RR) scales and posterior mean for estimate for the risk difference (RD); CrI: credible interval.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107586.

References

Albert, J.M., Wang, W., 2015. Sensitivity analyses for parametric causal mediation effect estimation. Biostatistics 16, 339–351. https://doi.org/10.1093/biostatistics/kxu048. arXiv: https://academic.oup.com/biostatistics/article-pdf/16/2/339/17741689/kxu048.pdf.

Barfield, R., Shen, J., Just, A.C., Vokonas, P.S., Schwartz, J., Baccarelli, A.A., VanderWeele, T.J., Lin, X., 2017. Testing for the indirect effect under the null for genome-wide mediation analyses. Genet. Epidemiol. 41, 824–833. https://doi.org/10.1002/gepi.22084. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22084.

Caubet, M., Samoilenko, M., Lefebvre, G., 2022. ExactMed: exact mediation analysis for binary outcomes. https://CRAN.R-project.org/package=ExactMed. r package version 0.1.0.

Caubet-Fernandez, M., Drouin, S., Samoilenko, M., Morel, S., Krajinovic, M., Laverdière, C., Sinnett, D., Levy, E., Marcil, V., Lefebvre, G., 2019. A Bayesian multivariate latent t-regression model for assessing the association between corticosteroid and cranial radiation exposures and cardiometabolic complications in survivors of childhood acute lymphoblastic leukemia: a PETALE study. BMC Med. Res. Methodol. 19. URL. https://doi.org/10.1186/s12874-019-0725-9.

Centers for Disease Control and Prevention, National diabetes statistics report, https://www.cdc.gov/diabetes/data/statistics-report. (Accessed 25 January 2021).

Cole, S.R., Platt, R.W., Schisterman, E.F., Chu, H., Westreich, D., Richardson, D., Poole, C., 2009. Illustrating bias due to conditioning on a collider. Int. J. Epidemiol. 39, 417–420. https://doi.org/10.1093/ije/dyp334. arXiv: https://academic.oup.com/ije/article-pdf/39/2/417/18480441/dyp334.pdf.

Daniels, M.J., Roy, J.A., Kim, C., Hogan, J.W., Perri, M.G., 2012. Bayesian inference for the causal effect of mediation. Biometrics 68, 1028–1036. https://doi.org/10.1111/j.1541-0420.2012.01781.x. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2012.01781.x.

Doretti, M., Raggi, M., Stanghellini, E., 2022. Exact parametric causal mediation analysis for a binary outcome with a binary mediator. Stat. Methods Appl. 31, 87–108.

Elliott, M.R., Raghunathan, T.E., Li, Y., 2010. Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. Biostatistics 11, 353–372. https://doi.org/10.1093/biostatistics/kxp060. arXiv: https://academic.oup.com/biostatistics/article-pdf/11/2/353/17738196/kxp060.pdf.

Enders, C.K., Fairchild, A.J., MacKinnon, D.P., 2013. A Bayesian approach for estimating mediation effects with missing data. Multivar. Behav. Res. 48, 340–369. https://doi.org/10.1080/00273171.2013.784862. PMID: 24039298.

Galharret, J.-M., Philippe, A., 2019. Priors comparison in Bayesian mediation framework with binary outcome. HAL, https://hal.archives-ouvertes.fr/hal-02070053.

Gaynor, S.M., Schwartz, J., Lin, X., 2019. Mediation analysis for common binary outcomes. Stat. Med. 38, 512–529. https://doi.org/10.1002/sim.7945. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7945.

Hund, L., Bedrick, E.J., Miller, C., Huerta, G., Nez, T., Ramone, S., Shuey, C., Cajero, M., Lewis, J., 2015. A Bayesian framework for estimating disease risk due to exposure to uranium mine and mill waste on the navajo nation. J. R. Stat. Soc., Ser. A, Stat. Soc. 178, 1069–1091. http://www.jstor.org/stable/43965784.

Imai, K., Keele, L., Tingley, D., 2010a. A general approach to causal mediation analysis. Psychol. Methods 15, 309–334. http://imai.princeton.edu/research/BaronKenny.html.

Imai, K., Keele, L., Yamamoto, T., 2010b. Identification, inference, and sensitivity analysis for causal mediation effects. Stat. Sci. 25, 51–71. http://imai.princeton.edu/research/mediation.html.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572. https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

Kim, C., Daniels, M., Li, Y., Milbury, K., Cohen, L., 2018. A bayesian semiparametric latent variable approach to causal mediation. Stat. Med. 37, 1149–1161. https://doi.org/10.1002/sim.7572. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7572.

Kim, C., Daniels, M.J., Marcus, B.H., Roy, J.A., 2017. A framework for Bayesian nonparametric inference for causal effects of mediation. Biometrics 73, 401–409. https://doi.org/10.1111/biom.12575. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12575. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12575.

Lange, T., Vansteelandt, S., Bekaert, M., 2012. A simple unified approach for estimating natural direct and indirect effects. Am. J. Epidemiol. 176, 190–195. https://doi.org/10.1093/aje/kwr525. arXiv: https://academic.oup.com/aje/article-pdf/176/3/190/17339781/kwr525.pdf.

Lindmark, A., de Luna, X., Eriksson, M., 2018. Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. Stat. Med. 37, 1744–1762. https://doi.org/10.1002/sim.7620. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7620.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A.A., Lin, X., 2021. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. J. Am. Stat. Assoc., 1–15. https://doi.org/10.1080/01621459.2021.1914634.

Loeys, T., Moerkerke, B., Smet, O.D., Buysse, A., Steen, J., Vansteelandt, S., 2013. Flexible mediation analysis in the presence of nonlinear relations: beyond the mediation formula. Multivar. Behav. Res. 48, 871–894. https://doi.org/10.1080/00273171.2013.832132.

MacKinnon, D.P., Lockwood, C.M., Williams, J., 2004. Confidence limits for the indirect effect: distribution of the product and resampling methods. Multivar. Behav. Res. 39, 99–128. https://doi.org/10.1207/s15327906mbr3901_4. PMID: 20157642.

Marcoux, S., Drouin, S., Laverdière, C., Alos, N., Andelfinger, G.U., Bertout, L., Curnier, D., Friedrich, M.G., Kritikou, E.A., Lefebvre, G., Levy, E., Lippé, S., Marcil, V., Raboisson, M.-J., Rauch, F., Robaey, P., Samoilenko, M., Séguin, C., Sultan, S., Krajinovic, M., Sinnett, D., 2017. The PETALE study: late adverse effects and biomarkers in childhood acute lymphoblastic leukemia survivors. Pediatric Blood & Cancer 64, e26361. https://doi.org/10.1002/pbc.26361. https://onlinelibrary.wiley.com/doi/abs/10.1002/pbc.26361. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pbc.26361.

Miočević, M., Gonzalez, O., Valente, M.J., MacKinnon, D.P., 2018. A tutorial in Bayesian potential outcomes mediation analysis. Struct. Equ. Model. 25, 121–136. https://doi.org/10.1080/10705511.2017.1342541. PMID: 29910595.

Miočević, M., MacKinnon, D.P., Levy, R., 2017. Power in Bayesian mediation analysis for small sample research. Struct. Equ. Model. 24, 666–683. https://doi.org/10.1080/10705511.2017.1312407. PMID: 29662296.

Muller, D., Judd, C.M., Yzerbyt, V.Y., 2005. When moderation is mediated and mediation is moderated. J. Pers. Soc. Psychol. 89, 852–863. https://doi.org/10.1037/0022-3514.89.6.852.

Nguyen, T.Q., Schmid, I., Ogburn, E.L., Stuart, E.A. Clarifying causal mediation analysis for the applied researcher: effect identification via three assumptions and five potential outcomes. Methodology. https://doi.org/10.48550/arXiv.2011.09537.

Nuijten, M.B., Wetzels, R., Matzke, D., Dolan, C.V., Wagenmakers, E.-J., 2015. A default Bayesian hypothesis test for mediation. Behav. Res. Methods 47, 85–97. https://doi.org/10.3758/s13428-014-0470-2.

O'Brien, S.M., Dunson, D.B., 2004. Bayesian multivariate logistic regression. Biometrics 60, 739–746. https://doi.org/10.1111/j.0006-341X.2004.00224.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2004.00224.x. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2004.00224.x.

Park, S., Kaplan, D., 2015. Bayesian causal mediation analysis for group randomized designs with homogeneous and heterogeneous effects: simulation and case study. Multivar. Behav. Res. 50, 316–333. https://doi.org/10.1080/00273171.2014.1003770. PMID: 26610032.

Pearl, J., 2001. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence UAI'01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 411–420.

Pearl, J., 2012. The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prev. Sci. 13, 426–436. https://doi.org/10.1007/s11121-011-0270-1.

Polley, E., LeDell, E., Kennedy, C., Lendle, S., van der Laan, M., 2021. SuperLearner: super learner prediction. https://CRAN.R-project.org/package=SuperLearner. r package version 2.0-28.

Rijnhart, J.J., Twisk, J.W., Chinapaw, M.J., de Boer, M.R., Heymans, M.W., 2017. Comparison of methods for the analysis of relatively simple mediation models. Contemp. Clin. Trials Commun. 7, 130–135. https://doi.org/10.1016/j.conctc.2017.06.005. https://www.sciencedirect.com/science/article/pii/S2451865417300285.

Rijnhart, J.J., Valente, M.J., MacKinnon, D.P., Twisk, J.W., Heymans, M.W., 2021. The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction. Struct. Equ. Model. 28, 345–355. https://doi.org/10.1080/10705511.2020.1811709.

Robin, J.M., Greenland, S., 1992. Identifiability and exchangeability for direct and indirect effects. Epidemiology 3, 143–155. https://doi.org/10.1097/00001648-199203000-00013.

Samoilenko, M., Blais, L., Lefebvre, G., 2018. Comparing logistic and log-binomial models for causal mediation analyses of binary mediators and rare binary outcomes: evidence to support cross-checking of mediation results in practice. Observat. Stud. 4, 193–216.

Samoilenko, M., Lefebvre, G., 2018. Point: risk ratio equations for natural direct and indirect effects in causal mediation analysis of a binary mediator and a binary outcome—a fresh look at the formulas. Am. J. Epidemiol. 188, 1201–1203. https://doi.org/10.1093/aje/kwy275. arXiv: https://academic.oup.com/aje/article-pdf/188/7/1201/28890378/kwy275.pdf.

Samoilenko, M., Lefebvre, G., 2021. Parametric-regression–based causal mediation analysis of binary outcomes and binary mediators: moving beyond the rareness or commonness of the outcome. Am. J. Epidemiol. 190, 1846–1858. https://doi.org/10.1093/aje/kwab055. arXiv: https://academic.oup.com/aje/article-pdf/190/9/1846/40828444/kwab055.pdf.

Song, Y., Zhou, X., Kang, J., Aung, M.T., Zhang, M., Zhao, W., Needham, B.L., Kardia, S.L.R., Liu, Y., Meeker, J.D., Smith, J.A., Mukherjee, B., 2021. Bayesian sparse mediation analysis with targeted penalization of natural indirect effects. J. R. Stat. Soc., Ser. C, Appl. Stat. 70, 1391–1412. https://doi.org/10.1111/rssc.12518. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12518. arXiv: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12518.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S.L.R., Roux, A.V.D., Needham, B.L., Smith, J.A., Mukherjee, B., 2020. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. Biometrics 76, 700–710. https://doi.org/10.1111/biom.13189. https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13189. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13189.

Steen, J., Loeys, T., Moerkerke, B., Vansteelandt, S., 2017. Medflex: an r package for flexible mediation analysis using natural effect models. J. Stat. Softw. 76, 1–46. https://doi.org/10.18637/jss.v076.i11. https://www.jstatsoft.org/index.php/jss/article/view/v076i11.

Tchetgen, E.J.T., Shpitser, I., 2012. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis. Ann. Stat. 40, 1816–1845. https://doi.org/10.1214/12-AOS990.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K., 2014. Mediation: R package for causal mediation analysis. J. Stat. Softw. 59, 1–38. http://www.jstatsoft.org/v59/i05/.

Valeri, L., VanderWeele, T.J., 2013. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. psychological methods. Psychol. Methods 18, 137–155. https://doi.org/10.1037/a0031034.

VanderWeele, T.J., Vansteelandt, S., 2009. Conceptual issues concerning mediation, interventions and composition. Stat. Interface 2, 457–468.

VanderWeele, T.J., 2015. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, New York, NY.

VanderWeele, T.J., Valeri, L., Ananth, C.V., 2018. Counterpoint: mediation formulas with binary mediators and outcomes and the "Rare Outcome Assumption". Am. J. Epidemiol. 188, 1204–1205. https://doi.org/10.1093/aje/kwy281. arXiv: https://academic.oup.com/aje/article-pdf/188/7/1204/28890402/kwy281.pdf.

Wagner, B.D., Kroehl, M., Gan, R., Mikulich-Gilbertson, S.K., Sagel, S.D., Riggs, P.D., Brown, T., Snell-Bergeon, J., Zerbe, G.O., 2018. A multivariate generalized linear model approach to mediation analysis and application of confidence ellipses. Stat. Biosci. 10, 139–159. https://EconPapers.repec.org/RePEc:spr:stabio:v:10:y:2018:i:1:d:10.1007_s12561-017-9191-2.

Wang, C., Hu, J., Blaser, M.J., Li, H., 2019. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics 36, 347–355. https://doi.org/10.1093/bioinformatics/btz565. arXiv: https://academic.oup.com/bioinformatics/article-pdf/36/2/347/36208723/btz565.pdf.

Wang, K., 2020. Direct effect and indirect effect on an outcome under nonlinear modeling. Int. J. Biostat. 16, 20190158. https://doi.org/10.1515/ijb-2019-0158.

Wilson, C.L., Liu, W., Yang, J.J., Kang, G., Ojha, R.P., Neale, G.A., Srivastava, D.K., Gurney, J.G., Hudson, M.M., Robison, L.L., Ness, K.K., 2015. Genetic and clinical factors associated with obesity among adult survivors of childhood cancer: a report from the st. jude lifetime cohort. Cancer 121, 2262–2270. https://doi.org/10.1002/cncr.29153. https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.29153. arXiv: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.29153.

Yuan, Y., MacKinnon, D.P., 2009. Bayesian mediation analysis. Psychol. Methods 14, 301–322. https://doi.org/10.1037/a0016972.