Received: 24 March 2023 | Revised: 8 October 2023 | Accepted: 11 November 2023

DOI: 10.1002/bimj.202300089

Biometrical Journal

RESEARCH ARTICLE

Mediation analysis with case–control sampling: Identification and estimation in the presence of a binary mediator

Marco Doretti[1] | **Minna Genbäck**[2] | **Elena Stanghellini**[2,3]

[1]Department of Statistics, Computer Science, and Applications, University of Florence, Florence, Italy

[2]Department of Statistics, USBE, Umeå University, Umeå, Sweden

[3]Department of Economics, University of Perugia, Perugia, Italy

Correspondence
Marco Doretti, Department of Statistics, Computer Science, and Applications, University of Florence, viale Morgagni 59, 50134 Florence, Italy.
Email: marco.doretti@unifi.it

Funding information
Swedish Research Council, Grant/Award Number: 2019-01064

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

With reference to a stratified case–control (CC) procedure based on a binary variable of primary interest, we derive the expression of the distortion induced by the sampling design on the parameters of the logistic model of a secondary variable. This is particularly relevant when performing mediation analysis (possibly in a causal framework) with stratified case–control (SCC) data in settings where both the outcome and the mediator are binary. Despite being designed for parametric identification, our strategy is general and can be used also in a nonparametric context. With reference to parametric estimation, we derive the maximum likelihood (ML) estimator and the M-estimator of the joint outcome–mediator parameter vector. We then conduct a simulation study focusing on the main causal mediation quantities (i.e., natural effects) and comparing M- and ML estimation to existing methods, based on weighting. As an illustrative example, we reanalyze a German CC data set in order to investigate whether the effect of reduced immunocompetency on listeriosis onset is mediated by the intake of gastric acid suppressors.

KEYWORDS
collider node, distortion, logistic regression, odds ratio, secondary outcome

1 | INTRODUCTION

Retrospective sampling schemes are common practice in studies when either the outcome of interest is rare or some covariates are difficult to measure. In this context, it is too expensive or too lengthy to random sample the whole population. Instead, more efficiently, random samples are extracted from a partition of the population according to the outcome and possibly other stratifying factors. When the outcome is binary, this procedure is known as (stratified) case–control (CC) sampling design.

Though the implications of the retrospective sampling are widely understood when the only interest is the relationship between the covariates and the outcome (Breslow, 1996), less is known when the relationship between some of the covariates is also the object of inference. Typically, this situation arises in mediation analysis, where the aim is to decompose,

in a counterfactual framework, the total effect (TE) of a treatment/exposure (henceforth, exposure) on the outcome into a direct and an indirect effect, the second one due to a mediator. A mediator is a variable that is a response of the exposure and in turn influences the outcome (Pearl, 2001; VanderWeele, 2015). With reference to parametric settings, two equations are of interest, the first linking the response to the exposure and the mediator, the second linking the mediator to the exposure. If the probability of a unit to be in the sample depends on the outcome (and possibly some other stratifying factors), the relationship between the exposure and the mediator can be distorted and standard mediation methods fail.

To address this problem, VanderWeele and Vansteelandt (2010) propose to estimate the mediator equation on the subsample of controls. Under the rare outcome assumption, this can be approximately considered a random sample from the whole population (VanderWeele & Tchetgen Tchetgen, 2016). To embed information from cases, a unified likelihood approach has also been introduced, again assuming the outcome is rare in the population (Satten et al., 2022). Another strategy beyond the rare outcome assumption (also mentioned in VanderWeele & Vansteelandt 2010), is based on weighting, a proposal originally contained in some papers by Manski and coauthors (Manski & Lerman, 1977; Manski & McFadden, 1981) in the context of *choice-based* samples, which are the analogous of stratified case–control (SCC) samples within the econometric literature (see also van der Laan, 2008). Typically, this strategy requires the outcome prevalence (possibly within strata) to be known.

With reference to a binary mediator and a binary outcome, both modeled via logistic regression, we here propose a parametric approach to mediation analysis when the sampling scheme depends on the outcome, and possibly other stratifying covariates. We exploit knowledge provided by directed acyclic graphs (DAGs, Lauritzen 1996), where an additional binary node is representing the sampling scheme. In this regard, we are in line with Didelez, Kreiner et al. (2010). We build on the existing literature on parametric mediation analysis for a binary mediator and a binary outcome (Doretti et al., 2022; Stanghellini & Doretti, 2019) to derive the explicit expression of the distortion induced by the SCC sampling design on the mediator logistic equation, in line with recent results on the distortion induced by the sampling scheme on the linear regression parameters (Kartsonaki & Cox, 2023). We then embed this expression into two estimating procedures based on maximum likelihood (ML) and M-estimation. Our strategy can be easily adapted to achieve nonparametric identification and, in turn, to perform nonparametric estimation.

As the result is general and does not hinge on the rare outcome assumption, there are a number of additional situations where this approach can be of interest. The first one concerns informative missingness in logistic regression. It turns out that our work extends the derivations in Wang et al. (2017) to the continuous exposure context, also proposing two likelihood-based estimating procedures that can be easily implemented with standard statistical software. Another situation of interest arises in CC studies with secondary or additional outcomes (Richardson et al., 2007), which is particularly common in genetic epidemiology, where information on a secondary phenotype is also of interest (Wang & Shete, 2011). Similarly, it can be of use to correct for selection bias, provided that units can be seen as randomly sampled according to some outcome-related binary partition of the population. More generally, properly recovering the distortion induced on associations by conditioning on a collider node is fundamental for many other purposes. In this sense, this work extends to a parametric framework the results on odds ratios of Nguyen et al. (2019).

The remainder of the paper is organized as follows. In Section 2.1, we recall the main concepts of causal mediation within the counterfactual framework, including the notions of natural (in)direct effect as well as the assumptions needed to identify such effects from observable data, both in a nonparametric and a parametric framework. The impact of outcome-dependent sampling is formalized in Section 2.2. In Section 3 (and the related Appendix), our adjusting identification strategy is first presented within the parametric framework, and then extended to the nonparametric one. The specifics of the two parametric estimation methods considered (M-estimation and ML) are described in Section 4. In Section 5, we apply the proposed methodology to data coming from a German CC study on listeriosis. Our aim is to decompose, in a counterfactual framework, the effect of reduced immunocompetency (RI) on listeriosis into a direct and an indirect effect, with the latter due to gastric acid suppressor (GAS) intake. Section 6 reports evidence from a simulation study conducted to compare M- and ML estimators to weighting estimators, while some concluding remarks are given in Section 7.

2 | BACKGROUND

2.1 | Causal mediation analysis

Let A denote a generic exposure, while M and Y represent the binary mediator and the binary outcome, respectively. Also, let C be a set of observed covariates. To deal with causal mediation, we rely on the counterfactual framework and

denote by $Y(a)$ and $M(a)$ the random variables representing the outcome and the mediator under an intervention setting, possibly contrary to the facts, A to the level a. Also, we let $Y(a,m)$ represent the outcome when A and M are set to a and m, respectively. This notation allows to introduce nested counterfactuals. In detail, $Y(a,M(a))$ denotes the random outcome under an intervention setting A to a and leaving M to the (random) value it would naturally attain under the same exposure level a. Because of this interpretation, the equality $Y(a) = Y(a,M(a))$ is assumed to hold true, being typically referred to as *composition* assumption. On the other hand, $Y(a,M(a^\star))$ denotes the outcome under an intervention setting $A = a$ but leaving M to the value it would naturally attain after setting the exposure to another level, that is, $a^\star$.

In most cases, the casual TE on the outcome of shifting the level of A from $a^\star$ to a (conditional on $C = c$) is defined on the difference scale as $E(Y(a) - Y(a^\star) \mid C = c)$, and an additive decomposition is introduced where TE equals (invoking the composition assumption) the sum of two components: the natural direct effect (NDE) $E(Y(a,M(a^\star)) - Y(a^\star,M(a^\star)) \mid C = c)$ and the natural indirect effect (NIE) $E(Y(a,M(a)) - Y(a,M(a^\star)) \mid C = c)$ (Pearl 2001; see also Robins & Greenland 1992). When the outcome is binary, effects on other scales are also considered. Specifically, VanderWeele and coauthors introduce the notions of odds-ratio NDE

$$\mathrm{OR}^{\mathrm{NDE}}_{a,a^\star|c} = \frac{P(Y(a,M(a^\star)) = 1 \mid C = c)/P(Y(a,M(a^\star)) = 0 \mid C = c)}{P(Y(a^\star,M(a^\star)) = 1 \mid C = c)/P(Y(a^\star,M(a^\star)) = 0 \mid C = c)} \tag{1}$$

and NIE

$$\mathrm{OR}^{\mathrm{NIE}}_{a,a^\star|c} = \frac{P(Y(a,M(a)) = 1 \mid C = c)/P(Y(a,M(a)) = 0 \mid C = c)}{P(Y(a,M(a^\star)) = 1 \mid C = c)/P(Y(a,M(a^\star)) = 0 \mid C = c)} \tag{2}$$

(Valeri & VanderWeele, 2013; VanderWeele & Vansteelandt, 2010). Multiplication of the above effects returns (again, under composition) the odds-ratio TE

$$\mathrm{OR}^{\mathrm{TE}}_{a,a^\star|c} = \frac{P(Y(a) = 1 \mid C = c)/P(Y(a) = 0 \mid C = c)}{P(Y(a^\star) = 1 \mid C = c)/P(Y(a^\star) = 0 \mid C = c)}.$$

Clearly, such a decomposition becomes again additive on the log odds-ratio scale.

Regardless of the chosen scale, a number of assumptions are required in order to obtain *observed-data* expressions for NDEs and NIEs, that is, to identify these quantities from observed data. In detail, a sufficient set of assumptions includes *consistency*, *positivity* as well as more context-specific assumptions concerning confounding. In words, consistency postulates that counterfactual variables are equal to observable variables whenever conditioning on the appropriate levels of the exposure and of the mediator occurs. Formally, for every level a we have that, given $A = a$, $M(a) = M$, and $Y(a) = Y$. Moreover, for every pair (a,m) we have $Y(a,m) = Y$ whenever $A = a$ and $M = m$ (VanderWeele, 2009; VanderWeele & Vansteelandt, 2009). Positivity assumes that, for every (a,m,c) configuration, $P(A = a \mid C = c)$ and $P(M = m \mid A = a, C = c)$ are strictly greater than 0 (Imai et al., 2010). Clearly, for continuous A the former probability is replaced by the corresponding density function. As for confounding, it is assumed that, conditional on covariates, no residual confounding exists for the relationship between (i) the exposure and the outcome, (ii) the mediator and the outcome, and (iii) the exposure and the mediator. In the language of conditional independence (Dawid, 1979), these assumptions can be formalized as (i) $Y(a,m) \perp\!\!\!\perp A \mid C$, (ii) $Y(a,m) \perp\!\!\!\perp M \mid A = a, C$, and (iii) $M(a) \perp\!\!\!\perp A \mid C$. The assumption completing the sufficient set is the so-called (iv) *cross-world independence* assumption $Y(a,m) \perp\!\!\!\perp M(a^\star) \mid C$, which entails a conditional independence between counterfactual variables referring to two different "worlds," that is, interventional settings where A is fixed to a and $a^\star$.

In an experimental context, (i) and (iii) could be enforced by randomizing A, possibly within the levels of C (Tchetgen Tchetgen, 2013). However, randomization of the treatment would not automatically remove confounding for the mediator–outcome relationship. In general, a sufficiently rich set of covariates C should be collected for (i)–(iii) to hold. As for the cross-world independence assumption (iv), the simultaneous involvement of two different interventional settings implies that no randomized experiment can be designed to enforce it (Tchetgen Tchetgen & VanderWeele, 2014), a fact complicating interpretation in observational studies, too. Notice that, without additional assumptions, a recanting witness (that is, a mediator-outcome confounder that is also affected by the exposure; see Avin et al. 2005) will contradict assumption (iv), even when all variables are measured (Andrews and Didelez, 2021). When relying on (iv), investigators need to be very cautious, as these are common structures.

Under the above assumptions, natural direct and indirect effects can be nonparametrically identified from observed data. In particular, an *observed-data* expression for $P(Y(a, M(a^\star)) = y)$ can be obtained, for $y \in \{0, 1\}$ and for every $(a, a^\star)$ pair, with Pearl's mediation formula (Pearl, 2001, 2010) as

$$P(Y(a, M(a^\star)) = y \mid c) = \sum_m P(Y = y \mid a, m, c)P(M = m \mid a^\star, c), \tag{3}$$

where $P(\cdot \mid k)$ is a shorthand for $P(\cdot \mid K = k)$ used when ambiguities do not occur. This results enables nonparametric identification of the effects in (1) and (2) and, in turn, of the corresponding TE.

In a parametric context, logistic regression models are often postulated for both the outcome and the mediator. In particular, we let

$$\text{logit}\{P(Y = 1 \mid a, c, m)\} = \boldsymbol{x}_y^\top \boldsymbol{\beta} \tag{4}$$

and

$$\text{logit}\{P(M = 1 \mid a, c)\} = \boldsymbol{x}_m^\top \boldsymbol{\delta}, \tag{5}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are coefficient vectors, while $\boldsymbol{x}_y$ and $\boldsymbol{x}_m$ contain the values of the explanatory variables (including the constant term). Such a general formulation allows for the presence, in both models, of first- and higher-order (e.g., quadratic) effects as well as of interaction terms. In this setting, starting from (3) parametric expressions for the effects in (1) and (2) have been derived. Specifically, we here refer to the approach in Doretti et al. (2022), that extend previous work of Valeri and VanderWeele (2013) to settings where the exposure (and the mediator) interact with the covariates (also outside the rare outcome case). The resulting formulas are reported in Section 1 of the Supplementary material.

2.2 | Outcome-dependent sampling

As is well-known, in the presence of outcome-dependent sampling some data distributions are artificially altered, thereby introducing some degree of distortion. Like in other related approaches (see, e.g., Didelez, Kreiner et al. 2010), we formalize this fact by introducing a selection indicator variable, W, such that data are available just on the selected units, that is, those with $W = 1$. A straightforward extension, that is not here considered, is when data can be seen as a random sample extracted from the population with $W = 1$.

The binary variable W is influenced by Y via a known probabilistic mechanism, that can be either marginal or conditional on the strata formed by a set of categorical background covariates. These covariates, denoted by B, might affect A, M, and Y as well, possibly acting as confounders. When B is empty, this scheme corresponds to the unconditional CC design. Otherwise, the SCC design arises. We also indicate by Z an additional set of covariates (of any nature) that do not influence the selection mechanism W, but that, together with B, might be needed to address confounding. Thus, we have $C = (B, Z)$. The overall framework is represented in the two DAGs in Figure 1, where the W node is squared in order to pinpoint its nature of selection node.

In what follows, we account for the presence of B and thus refer to the SCC setting, with the CC being a special case thereof. In order to ease the exposition, without loss of generality we think of B and Z as of singleton variables, with B taking values in $\{1, \dots, N_B\}$ to represent the N_B different strata formed by background variables. Thus, the $\boldsymbol{x}_y$ and $\boldsymbol{x}_m$ vectors in (4) and (5) contain, together with the values of other explanatory variables, stratum-specific indicator variables $\mathbb{I}\{B = b\}$ for $b = 2, \dots, N_B$ (the usual corner-point parameterization taking the first stratum as reference is adopted).

The impact of SCC sampling can be expressed in the language of conditional independence by stating that Y and M are not conditionally independent of W given the respective parent node sets, with the obvious consequence that

$$\begin{aligned} logit\{P(Y = 1 \mid a, b, z, m, W = 1)\} &\neq \boldsymbol{x}_y^\top \boldsymbol{\beta} \\ logit\{P(M = 1 \mid a, b, z, W = 1)\} &\neq \boldsymbol{x}_m^\top \boldsymbol{\delta}, \end{aligned}$$

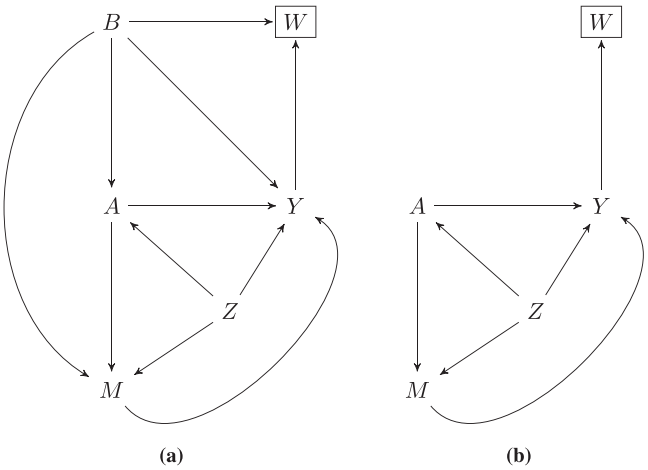

FIGURE 1 Mediation scheme for the (A, M, Y) triplet with covariates in a (a) stratified and (b) unconditional case–control sampling setting. Only categorical variables B affect W, whereas both B and Z may act as confounders by influencing (some of) the variables in (A, M, Y).

which shows that the models postulated in the right-hand sides of (4) and (5) are not identified in the subpopulation of selected units, and that adjusting approaches are needed. However, by inspection of the DAG it is possible to see that

$$M \perp\!\!\!\perp W \mid (A, B, Z, Y) \tag{6}$$

and therefore Y can be seen as a *selection bias breaking* variable, a term introduced elsewhere in the literature (Geneletti et al., 2009). This property, together with the explicit expressions of the distortion induced by the sampling design on the parameters of (4) and (5), forms the basis of our proposal. Notice that the minimum conditioning set for independence of M and W to hold is (B, Y). However, since we are interested in the mediating role of M on the pathway from A to Y, all variables in the conditioning set of (6) are relevant.

3 | IDENTIFICATION

In order to identify the coefficients of model (4) from SCC data, it is well-known that simple coefficient adjustments involving the conditional population prevalences $\pi_b = P(Y = 1 \mid B = b)$ suffice (Breslow et al., 1988; Fears & Brown, 1986; Prentice & Pyke, 1979). In detail, the model expression holding in the subpopulation of selected units is given by

$$\text{logit}\{P(Y = 1 \mid a, b, z, m, W = 1)\} = \boldsymbol{x}_y^\top \boldsymbol{\beta}^\star, \tag{7}$$

where $\boldsymbol{\beta}^\star$ is equal to $\boldsymbol{\beta}$ except for modifications concerning the intercept (β_{INT}) and the coefficient subvector $\boldsymbol{\beta}_{\text{B}} = (\beta_2, \dots, \beta_{N_{\text{B}}})^\top$ related to the $N_{\text{B}} - 1$ stratum indicator variables. Specifically, letting $p_b = P(Y = 1 \mid W = 1, B = b)$ be the proportion of cases in the SCC sample for stratum b, the replacing elements in $\boldsymbol{\beta}^\star$ are

$$\begin{aligned} \beta^\star_{\text{INT}} &= \beta_{\text{INT}} + \log(k_1) \\ \boldsymbol{\beta}^\star_{\text{B}} &= \boldsymbol{\beta}_{\text{B}} + \log(\boldsymbol{k}_{-1}), \end{aligned} \tag{8}$$

where $\boldsymbol{k}_{-1} = k_1^{-1} \cdot (k_2, \dots, k_{N_{\text{B}}})^\top$ and, for $b \in \{1, \dots, N_{\text{B}}\}$,

$$\begin{aligned} k_b &= \frac{P(W = 1 \mid Y = 1, B = b)}{P(W = 1 \mid Y = 0, B = b)} \\ &= \frac{P(Y = 1 \mid W = 1, B = b)}{P(Y = 0 \mid W = 1, B = b)} \cdot \frac{P(Y = 0 \mid B = b)}{P(Y = 1 \mid B = b)} = \frac{p_b}{1 - p_b} \cdot \frac{1 - \pi_b}{\pi_b} \end{aligned}$$

(see Appendix A for details). Each of the correction terms above is known whenever π_b is. Indeed, $p_b/(1-p_b)$ is the stratum-specific CC ratio, which is fixed by design. We here assume π_b ($b = 1, \dots, N_B$) to be known from external auxiliary information sources.

We turn now to the model for $P(M = 1 \mid a, b, z, y, W = 1)$. From (6), we know that $P(M = 1 \mid a, b, z, y, W = 1) = P(M = 1 \mid a, b, z, y)$. After some derivations, see Appendix A, it is possible to see that

$$\text{logit}\{P(M = 1 \mid a, b, z, y)\} = \boldsymbol{x}_m^\top \boldsymbol{\delta} + o(y, \boldsymbol{x}_y; \boldsymbol{\beta}), \tag{9}$$

where

$$o(y, \boldsymbol{x}_y; \boldsymbol{\beta}) = \log \frac{P(Y = y \mid a, b, z, M = 1)}{P(Y = y \mid a, b, z, M = 0)} \tag{10}$$

is a correction term depending on y and the coefficient vector $\boldsymbol{\beta}$. The parametric expression of it is (A3) of Appendix A.

As mentioned in the Introduction, the proposed identification strategy can be implemented also in a nonparametric setting (Pearl, 2001, 2010), thereby enabling causal mediation via the nonparametric formula in (3). The key idea is formalized in the second part of Appendix A and relies on the postulated conditional independence structure, without invoking any functional form.

4 | PARAMETRIC ESTIMATION

Starting from the identification result in (7), inference on $\boldsymbol{\beta}^\star$ can be conducted from SCC data in a standard ML framework (Anderson, 1972; Prentice & Pyke, 1979). In parallel to this coefficient adjusting approach, weighting methods have also been proposed within the econometric literature dealing with *choice-based* samples (Manski & Lerman, 1977; Manski & McFadden, 1981), which can essentially be thought of as the analogous of SCC samples (Breslow, 1996). These methods consist in fitting the population model (4) to the SCC sample, assigning cases and controls in each stratum b a weight equal to π_b/p_b and $(1-\pi_b)/(1-p_b)$, respectively. The weighting estimators are consistent but less efficient than the ones obtained with parameter corrections if the regression model is correctly specified (King & Zeng, 2001). On the other hand, the parameter correction method might be less robust than weighting in the case of model misspecification (Xie & Manski, 1989).

Importantly, the weighting approach was also extended to model variables different from the original outcome (sometimes termed secondary or additional outcomes, Richardson et al. 2007); see, for example, the more general work by van der Laan (2008). Such an extension is of particular relevance in a mediation framework, where the population parameters of the mediator model also need to be recovered from SCC data in order to achieve effect decomposition (VanderWeele & Vansteelandt, 2010). In our setting, this would correspond to fitting the population model (5) to the SCC sample with the same weighting scheme as above.

Here, we introduce an alternative framework where estimation of the $\boldsymbol{\delta}$ vector from SCC data exploits the identification result in (9). Since $\boldsymbol{\beta}$ is also involved in that equation, our framework is designed for estimating the combined parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ altogether. This also fits with the mediation setting, where typically both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ have to be estimated at the same time. Two estimation strategies are discussed: in Section 4.1 we present an M-estimation approach, while in Section 4.2 the above-mentioned ML framework is extended to our bivariate context.

4.1 | M-estimation

In a formal M-estimation setting (Huber, 1964), the M-estimator $\hat{\boldsymbol{\theta}}_M = (\hat{\boldsymbol{\beta}}_M^\top, \hat{\boldsymbol{\delta}}_M^\top)^\top$ is the vector satisfying

$$\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_M) = \begin{pmatrix} \boldsymbol{s}_y(\hat{\boldsymbol{\theta}}_M) \\ \boldsymbol{s}_m(\hat{\boldsymbol{\theta}}_M) \end{pmatrix} = \mathbf{0}_{2d_\theta},$$

where d_θ is the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{s}_y(\cdot)$ and $\boldsymbol{s}_m(\cdot)$ are the observed score functions associated to the models in (7) and (9), respectively (see Appendix B for their expressions). In practice, a two-step procedure is required. First, the $\hat{\boldsymbol{\beta}}_M$ estimate is

obtained by fitting model (7) and implementing the corrections for β_{INT} and $\boldsymbol{\beta}_{\text{B}}$. Then, $\hat{\boldsymbol{\delta}}_{\text{M}}$ is computed by fitting model (9), where $o(y, \boldsymbol{x}_y; \hat{\boldsymbol{\beta}}_{\text{M}})$ is included as an offset term. Notice that $\hat{\boldsymbol{\delta}}_{\text{M}}$ is called a *partial* M-estimator (Stefanski & Boos, 2002), since it requires the $\hat{\boldsymbol{\beta}}_{\text{M}}$ estimate of the first step to be plugged-in in place of the unknown $\boldsymbol{\beta}$ vector (Randles, 1982). In this case, knowledge of the adjusting terms $\log(k_1)$ and $\log(\boldsymbol{k}_{-1})$ is essential, since the estimated offset $o(y, \boldsymbol{x}_y; \hat{\boldsymbol{\beta}}_{\text{M}})$ contains $\hat{\beta}_{\text{INT}}$ and $\hat{\boldsymbol{\beta}}_{\text{B}}$, rather than $\hat{\beta}^{\star}_{\text{INT}}$ and $\hat{\boldsymbol{\beta}}^{\star}_{\text{B}}$, in its expression.

The variance–covariance matrix of the $\hat{\boldsymbol{\theta}}_{\text{M}}$ estimator is given by

$$V(\hat{\boldsymbol{\theta}}_{\text{M}}) = \frac{1}{n} A(Y_i, \boldsymbol{\theta})^{-1} B(Y_i, \boldsymbol{\theta}) \{A(Y_i, \boldsymbol{\theta})^{-1}\}^{\top},$$

where

$$A(Y_i, \boldsymbol{\theta}) = E\left\{ \frac{\partial \boldsymbol{\Psi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}} \right\} \qquad B(Y_i, \boldsymbol{\theta}) = E\left\{ \boldsymbol{\Psi}_i(\boldsymbol{\theta}) \boldsymbol{\Psi}_i^{\top}(\boldsymbol{\theta}) \right\}$$

and $\boldsymbol{\Psi}_i(\boldsymbol{\theta})$ is the score random vector of a generic unit i. The finite-sample estimate of $V(\hat{\boldsymbol{\theta}}_{\text{M}})$ can be computed as

$$\hat{V}(\hat{\boldsymbol{\theta}}_{\text{M}}) = \frac{1}{n} A(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{\text{M}})^{-1} B(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{\text{M}}) \left\{ A(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{\text{M}})^{-1} \right\}^{\top}, \tag{11}$$

where $A(\boldsymbol{y}, \boldsymbol{\theta})$ and $B(\boldsymbol{y}, \boldsymbol{\theta})$ are the sample equivalent of $A(Y_i, \boldsymbol{\theta})$ and $B(Y_i, \boldsymbol{\theta})$ (Stefanski & Boos, 2002). Their expressions are also reported in Appendix B.

4.2 | ML estimation

Since we are dealing with an SCC sample, the likelihood of the observed data corresponds to the conditional density of the (M, A, Z) random vector given (Y, B) and $W = 1$. As not only the $\boldsymbol{\beta}$ parameters of (4) are the object of inference, but also the $\boldsymbol{\delta}$ parameters of (5), ML theory developed by Prentice and Pyke (1979) should be extended. With reference to discrete choice models, Imbens (1992) provides more general results that can be applied to this context. Specifically, given n independent sample units indexed by $i = 1, \ldots, n$, it follows from Imbens (1992) that ML estimation of $\boldsymbol{\theta}$ can be performed by maximizing

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(Y_i = y_i, M_i = m_i \mid a_i, z_i, b_i, W_i = 1). \tag{12}$$

An additional factorization of (12) leads to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(Y_i = y_i \mid a_i, z_i, b_i, W_i = 1) P(M_i = m_i \mid a_i, z_i, b_i, y_i, W_i = 1),$$

so that the corresponding log-likelihood $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ is equal to

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} & y_i \log p_{y^{\star}_{(m)},i} + (1 - y_i) \log \left\{ 1 - p_{y^{\star}_{(m)},i} \right\} \\ & + m_i \log p_{my,i} + (1 - m_i) \log \left\{ 1 - p_{my,i} \right\}, \end{aligned} \tag{13}$$

where $p_{y^{\star}_{(m)},i}$ and $p_{my,i}$ are shorthands for the two probabilities $P(Y_i = 1 \mid a_i, z_i, b_i, W_i = 1)$ and $P(M_i = 1 \mid a_i, z_i, b_i, y_i, W_i = 1)$, respectively. While the latter is related to the parameter vector $\boldsymbol{\theta}$ via (9), the former involves marginalization over the binary mediator M. The corresponding model on the logistic scale can be written as

$$\text{logit}\{P(Y = 1 \mid a, z, b, W = 1)\} = \text{logit}\{P(Y = 1 \mid a, z, b, M = 0, W = 1)\} + g(\boldsymbol{x}_y, \boldsymbol{x}_m; \boldsymbol{\beta}, \boldsymbol{\delta}), \tag{14}$$

where the first term is (7) evaluated at $M = 0$ while

$$g(\boldsymbol{x}_y, \boldsymbol{x}_m; \boldsymbol{\beta}, \boldsymbol{\delta}) = \log \frac{P(M = 0 \mid Y = 0, a, z, b, W = 1)}{P(M = 0 \mid Y = 1, a, z, b, W = 1)}$$

is a correction term which depends on $\boldsymbol{x}_y$, $\boldsymbol{x}_m$ as well as on both parameter vectors. See Appendix C for the parametric expression of (14).

Within this framework, the ML estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\text{ML}}$, can be obtained by maximizing the log-likelihood in (13) via the usual iterative methods. Since these methods are likely to suffer from local maxima problems, it is advisable to set several starting points fluctuating around a sensible deterministic choice (see Section 5.2 for an account of the approach undertaken within our specific application). To further enhance the performance of optimization algorithms, it is typically useful to provide the expression of the log-likelihood gradient $\boldsymbol{s}(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, which is reported in Appendix D.

In line with standard theory, the estimated variance–covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is given by $\{-H(\hat{\boldsymbol{\theta}}_{\text{ML}})\}^{-1}$, where $H(\hat{\boldsymbol{\theta}}_{\text{ML}}) = \{\partial \boldsymbol{s}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\top}\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{ML}}}$ is the Hessian matrix computed at $\hat{\boldsymbol{\theta}}_{\text{ML}}$. Alternatively, sandwich estimation (Royall, 1986; White, 1980) can be performed via

$$\widehat{\text{cov}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = \{-H(\hat{\boldsymbol{\theta}}_{\text{ML}})\}^{-1} \boldsymbol{Q}(\hat{\boldsymbol{\theta}}_{\text{ML}}) \{-H(\hat{\boldsymbol{\theta}}_{\text{ML}})\}^{-1},$$

where $\boldsymbol{Q}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = \left\{\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta}) \boldsymbol{s}_i(\boldsymbol{\theta})^{\top}\right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{ML}}}$ and $\boldsymbol{s}_i(\boldsymbol{\theta})$ is the individual contribution of the ith unit to $\boldsymbol{s}(\boldsymbol{\theta})$.

5 | CASE STUDY

5.1 | The listeriosis data set

To illustrate the proposed approach, we reconsider the data set analyzed by Preußel et al. (2015), where risk factors related to listeriosis in Germany are investigated. Listeriosis is an infection primarily contracted through the intake of food contaminated with *Listeria monocytogenes* bacteria. It mainly affects older adults, individuals with weakened immune systems, and pregnant women (Goulet et al., 2012). Consequences might be quite severe, ranging from the development of life-threatening conditions for the fetus—in the case of pregnancy—to severe illness or death in the other cases.

The study by Preußel et al. (2015) focusses on sporadic nonpregnancy-associated listeriosis. Many food-related risk factors, like consumption of cold cooked sausages and presliced cheese, are considered in combination with personal characteristics as well as other risk factors such as RI or the intake of GASs. The latter act as effect modifiers and can also be thought of as risk factors themselves when focus goes beyond food-related factors (Bavishi & Dupont, 2011; Goulet et al., 2012; Ho et al., 1986; Mook et al., 2013). In this application, we build upon this framework and extend it to a mediation setting in order to investigate whether—and to what extent—the additional vulnerability to listeriosis due to RI goes through the intake of GASs, which are commonly prescribed drugs for patients with RI (Ahrens et al., 2012; Preußel et al., 2015). In line with the original approach of Preußel et al. (2015), the exposure (RI) is coded as a categorical variable with three mutually exclusive levels: 0 = none, 1 = RI due to immunocompromising diseases without immunosuppressive therapies (e.g., diabetes, autoimmune disorder), and 2 = RI due to immunosuppressive therapies (e.g., chemotherapy, radiation therapy). The mediator (GAS intake) is binary (0 = no, 1 = yes). RI and GAS intake are considered present if occurring within 3 months from listeriosis onset (cases) or interview (controls).

As reported by Preußel et al. (2015), 732 sporadic (i.e., not due to an outbreak) nonpregnancy-associated listeriosis cases were ascertained in the time span from March 15, 2012 to December 31, 2013 (21.5 months) among the residents of German Federal states aged 40 or more (40+), with the exception of Bremen, which accounts for around 0.8% of the German population. Given that the 40+ German population on December 31, 2012 was 46,597,036,[1] the yearly incidence (accounting for the exclusion of Bremen) is given by $(12/21.5) \cdot \{732/(46,597,036 \cdot 0.992)\} = 8.84 \times 10^{-6}$. The overall listeriosis prevalence can be reconstructed from such an incidence assuming that 7 days is the average length of infection. The obtained prevalence rate is $(7/365) \cdot 8.84 \times 10^{-6} = 1.70 \times 10^{-7}$.

[1] See Table 12411-0005 of the Federal Statistical Office of Germany database. The table can be accessed from this link, following the path corresponding to its number sequence.

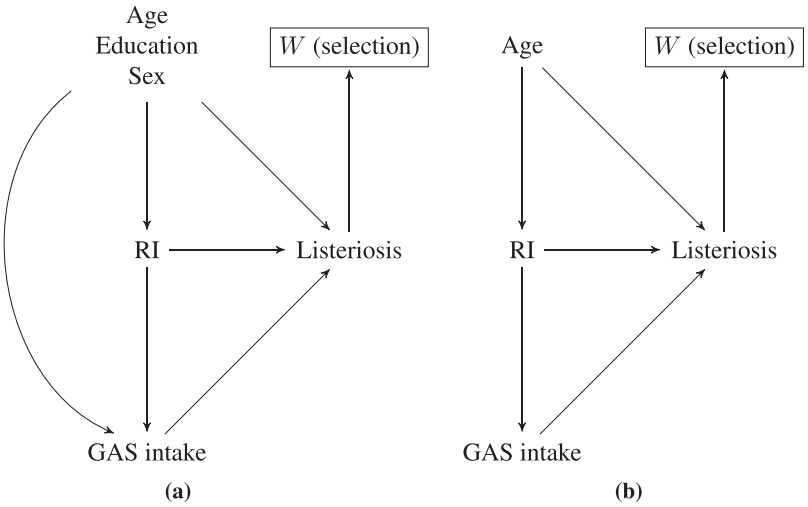

FIGURE 2 Directed acyclic graphs (DAGs) representing the (a) initially postulated and (b) final scheme when investigating the association between reduced immunocompetency (RI) and listeriosis through gastric acid suppressor (GAS).

In the study of Preußel et al. (2015), 1982 controls were enrolled along with cases from the population of subjects with no history of listeriosis which are accessible by telephone in Germany. Specifically, in order to obtain an SCC data set controls were sampled to get an equal number of individuals in the three age classes (ACs) 40–65, 66–75, and 76+ years, according to the age distribution of cases in the 2004–2011 years. However, only 109 of the 732 originally ascertained cases entered the study, and only 99 remained in the data set after preliminary data cleansing/missing data removal. The age distribution of the 109 cases entering the study (38.53% 40–65 years, 35.78% 66–75 years, and 25.69% 76+ years) differs from the uniform distribution of controls, and so does the distribution of the 99 cases in the final data set (39.40% 40–65 years, 36.36% 66–75 years, and 24.24% 76+ years). Because of these discrepancies, treating the data set as if it was gathered from an SCC design might not be entirely correct. Also, in an unconditional CC design the age distribution of the controls should be close to the age distribution of the 40+ population (65.94% 40–65 years, 18.83% 66–75 years, and 15.23% 76+ years on December 31, 2012), differing by sampling error only. As expected, this is not the case for the data set at hand, either.

We obviate the above-mentioned issues by randomly selecting a fixed number of controls (eight times the number of cases) in a way such that the age distribution of the German 40+ population is matched, thereby generating an unconditional CC sample. This sample, combined with the obtained prevalence rate, allows to implement the identification and estimation approach described in Sections 3 and 4. The resulting adjusting correction term for the intercept is $\log(k_1) = 13.511$. In what follows, estimation results from this CC sample are discussed, whereas in the Supplementary material (Section 2) estimates from an SCC sample (based on stratification on ACs) are reported.

5.2 | Results

The causal structure underlying the listeriosis data set is studied in Preußel et al. (2015), where evidence from the epidemiological literature is reported in justification of every postulated relationship between variables. Specifically, after marginalization over food-related factors, age, educational level, and gender are identified as exposure–outcome, mediator–outcome, and exposure–mediator confounders. In principle, this setting does not contradict assumption (iv), too. However, such an assumption may still be violated (see Section 2.1), so particular caution is needed since results will depend on its validity.

In accordance with the above data-generating process, we initially include age, educational level, and gender in both the outcome and mediator model (Figure 2a). However, subsequent analyses lead to keeping only age (coded with the three classes introduced above) in the outcome model, and no covariates in the mediator model (see Figure 2b). With this regard, it is worth to underline that model selection is conducted separately for the two models, within an M-estimation framework. In detail, selection procedures for the mediator model are implemented conditionally on the estimated offset term, $o(y, \boldsymbol{x}_{y_0}; \hat{\boldsymbol{\beta}}_{\mathrm{M}})$, resulting from the selected outcome model. Even at intermediate steps of the selection process, inference is based on the estimated variance–covariance matrix (11) rather than on the naive standard errors returned by statistical software.

Table 1 reports results for two pairs of logistic models fitted with M-estimation, ML, and weighting. Standard errors are obtained with a sandwich approach not only for M- and ML estimation (see Section 4), but also for weighting. ML

TABLE 1 Estimates (est.), standard errors (s.e.), and p-values (p) of logistic models for the outcome and the mediator without (top) and with (bottom) interactions between reduced immunocompetency (RI) and age class (AC) in the outcome model. AC = 1 is 40–65 years (baseline), AC = 2 is 66–75 years, and AC = 3 is 76+ years.

	M-estimation			ML			Weighting		
	est.	s.e.	p	est.	s.e.	p	est.	s.e	p
Outcome model									
Intercept	−16.963	0.228	0.000	−16.951	0.224	0.000	−17.113	0.245	0.000
RI = 1	1.318	0.287	0.000	1.309	0.287	0.000	1.416	0.296	0.000
RI = 2	2.209	0.278	0.000	2.199	0.279	0.000	2.177	0.280	0.000
AC = 2	1.016	0.263	0.000	1.018	0.261	0.000	0.973	0.283	0.001
AC = 3	0.747	0.305	0.014	0.729	0.307	0.018	0.725	0.311	0.020
GAS = 1	0.974	0.307	0.002	0.976	0.318	0.002	1.161	0.384	0.002
Mediator model									
Intercept	−3.279	0.219	0.000	−3.280	0.218	0.000	−3.159	0.208	0.000
RI = 1	0.870	0.333	0.009	0.869	0.335	0.009	0.529	0.421	0.209
RI = 2	0.855	0.344	0.013	0.854	0.351	0.015	0.594	0.472	0.208
Outcome model									
Intercept	−17.088	0.296	0.000	−17.103	0.298	0.000	−17.250	0.299	0.000
RI = 1	0.872	0.564	0.122	0.899	0.555	0.105	0.798	0.566	0.159
RI = 2	2.697	0.393	0.000	2.728	0.391	0.000	2.668	0.398	0.000
AC = 2	0.902	0.477	0.059	0.959	0.469	0.041	0.853	0.478	0.074
AC = 3	1.417	0.465	0.002	1.440	0.459	0.002	1.382	0.465	0.003
RI = 1, AC = 2	1.228	0.752	0.102	1.136	0.738	0.124	1.484	0.794	0.061
RI = 2, AC = 2	−0.575	0.656	0.381	−0.653	0.644	0.311	−0.574	0.669	0.391
RI = 1, AC = 3	−0.301	0.780	0.700	−0.330	0.776	0.670	−0.111	0.812	0.891
RI = 2, AC = 3	−1.432	0.674	0.034	−1.514	0.677	0.025	−1.255	0.703	0.074
GAS = 1	0.973	0.309	0.002	0.976	0.318	0.002	1.267	0.412	0.002
Mediator model									
Intercept	−3.279	0.219	0.000	−3.280	0.218	0.000	−3.159	0.208	0.000
RI = 1	0.870	0.330	0.008	0.869	0.335	0.009	0.529	0.421	0.209
RI = 2	0.856	0.347	0.014	0.854	0.351	0.015	0.594	0.472	0.208

Abbreviations: GAS, gastric acid suppressor; ML, maximum likelihood.

estimation is performed through direct maximization of the log-likelihood function by means of the `maxLik` function in `R` (Henningsen & Toomet, 2011). Multiple starting values (10) are obtained by perturbing the $\hat{\theta}_M$ vector obtained from the M-estimates with a random normal deviation with null mean and standard deviation equal to 0.5. All attempts converge to the same solution. However, unstable solutions might occur when starting values are determined totally at random.

In the first pair of models (top part of Table 1), the outcome equation does not contain interaction effects between AC and RI, whereas in the second pair (bottom part of the table) it does. Conversely, the mediator equation always includes RI effects only, meaning that its selection strategy is not influenced by whether or not interactions are included in the outcome model (parameter estimates are in practice not sensitive to this change, too). The choice between the outcome model with and without interactions is not straightforward. Indeed, only the coefficient for RI = 2, AC = 3 is significant. However, an overall ANOVA test comparing them (within the M-estimation framework) provides evidence in favor of the extended one (χ^2 deviance statistics equal to 10.787 with 4 degrees of freedom, p-value 0.029). Evidence is mixed also with regard to the Akaike and Bayesian Information Criteria, with the former favoring the interaction model (522.67 vs. 525.46 for the no interaction model), and the latter favoring the no interaction model (554.21 vs. 570.59 for the interaction model). For this reason, both pairs are kept for reference. In order to ease interpretation, the outcome regression coefficients of AC and RI are also linearly combined to obtain contrasts, on the log odds-ratio scale, between any pair of levels. These are reported, together with their standard errors, in Table 2, both for the no interaction (top part) and interaction (bottom part) setting. Clearly, in the latter case the estimated contrasts of each factor are conditional on the level of the other.

TABLE 2 Estimated log odds-ratio contrasts (with standard errors and *p*-values) between any pair of levels for outcome effects of age class (AC) and reduced immunocompetency (RI), both for the no interaction (top) and interaction (bottom) model. Stars denote contrasts related to regression coefficients, already in Table 1.

	M-estimation			ML			Weighting		
	est.	s.e.	*p*	est.	s.e.	*p*	est.	s.e	*p*
Outcome model without interactions									
	Age Class (AC)								
2 vs. 1⋆	1.016	0.263	0.000	1.018	0.261	0.000	0.973	0.283	0.001
3 vs. 1⋆	0.747	0.305	0.014	0.729	0.307	0.018	0.725	0.311	0.020
3 vs. 2	−0.268	0.322	0.405	−0.289	0.326	0.377	−0.249	0.331	0.452
	RI								
1 vs. 0⋆	1.318	0.287	0.000	1.309	0.287	0.000	1.416	0.296	0.000
2 vs. 0⋆	2.209	0.278	0.000	2.199	0.279	0.000	2.177	0.280	0.000
2 vs. 1	0.891	0.286	0.002	0.890	0.287	0.002	0.761	0.301	0.012
Outcome model with interactions									
	AC = 2 versus AC = 1								
RI = 0⋆	0.902	0.477	0.059	0.959	0.469	0.041	0.853	0.478	0.074
RI = 1	2.131	0.576	0.000	2.095	0.569	0.000	2.337	0.609	0.000
RI = 2	0.327	0.450	0.467	0.307	0.441	0.486	0.280	0.470	0.551
	AC = 3 versus AC = 1								
RI = 0⋆	1.417	0.465	0.002	1.440	0.459	0.002	1.382	0.465	0.003
RI = 1	1.117	0.624	0.074	1.110	0.626	0.076	1.271	0.655	0.053
RI = 2	−0.015	0.488	0.976	−0.073	0.498	0.883	0.127	0.519	0.807
	AC = 3 versus AC = 2								
RI = 0	0.515	0.520	0.322	0.481	0.509	0.345	0.528	0.517	0.307
RI = 1	−1.014	0.511	0.047	−0.985	0.533	0.065	−1.067	0.539	0.048
RI = 2	−0.342	0.550	0.534	−0.380	0.555	0.493	−0.153	0.587	0.794
	RI = 1 versus RI = 0								
AC = 1⋆	0.872	0.564	0.122	0.899	0.555	0.105	0.798	0.566	0.159
AC = 2	2.101	0.492	0.000	2.035	0.491	0.000	2.282	0.521	0.000
AC = 3	0.572	0.537	0.287	0.569	0.543	0.295	0.687	0.562	0.222
	RI = 2 versus RI = 0								
AC = 1⋆	2.697	0.393	0.000	2.728	0.391	0.000	2.668	0.398	0.000
AC = 2	2.122	0.523	0.000	2.076	0.513	0.000	2.094	0.530	0.000
AC = 3	1.265	0.546	0.021	1.214	0.551	0.028	1.413	0.570	0.013
	RI = 2 versus RI = 1								
AC = 1	1.825	0.545	0.001	1.829	0.537	0.001	1.870	0.547	0.001
AC = 2	0.021	0.485	0.965	0.041	0.485	0.933	−0.188	0.523	0.719
AC = 3	0.693	0.573	0.227	0.646	0.592	0.275	0.726	0.602	0.228

Abbreviation: ML, maximum likelihood.

Tables 1 and 2 show that ML and M-estimation provide very similar results for both model pairs. In particular, the estimated coefficient for the effect of GAS intake on listeriosis development is always positive (very close to 0.97, *p*-value 0.002). This confirms that GAS intake remains positively associated with listeriosis, even when controlling for RI (Preußel et al., 2015). Similar conclusions can be drawn for RI effects due to both diseases (RI = 1) and therapies (RI = 2) on GAS intake, with all coefficients being around 0.86 (*p*-values lower than 0.02). In light of these results, we can conclude that there is no evidence of a difference between diseases and therapies with respect to their effect on GAS intake.

With regard to effects of age on the outcome, the interaction model shows an increasing pattern for the no RI (RI = 0) group, although the shift from the 40–65 to the 66–75 AC is barely significant for M-estimation (*p*-value 0.059), and that

from the 66–75 to the 76+ class is not significant (p-values greater than 0.30). For RI = 1, a rather relevant positive difference is estimated for the contrast involving the second and the first ACs. Conversely, the estimated difference between the third and the first class is smaller (and barely significant, p-values around 0.07). As a consequence, the risk for listeriosis in the second class is higher than that of the third one, with the difference being significant to some extent (p-value 0.047 for M-estimation and 0.065 for ML). For the RI = 2 group, no significant differences among ACs are present. In the no interaction model, a unique pattern is estimated for all RI groups which is similar to the one described for the RI = 1 group in the model with interactions, although with reduced magnitudes and no significance for the contrast between the third and second AC (p-values greater than 0.35).

As for RI effects on the outcome, in the no interaction model we notice a significantly increased vulnerability to listeriosis when moving from RI = 0 to both RI = 1 and RI = 2 (p-values lower than 0.001). The difference between the latter levels is significant, too, with immunosuppressive therapies further enhancing the risk for listeriosis with respect to immunocompromising diseases (p-values 0.002). Effect magnitudes are in line with the findings in the original study (see Preußel et al., 2015). In the model with interactions, such a pattern of effect directions is preserved within each AC, though with different magnitudes and significance levels. Overall, we can conclude that: (i) only therapies have a significant effect in the 40–65 class, (ii) differences between therapies and diseases are essentially null in the 66–75 class, and (iii) RI effects considerably lessen in the 76+ class.

In terms of effect direction and statistical significance, results from the weighting approach are not extremely dissimilar to those previously reported. However, some noteworthy discrepancies emerge which concern GAS intake effects in the outcome model and RI effects in the mediator model. For the former, we observe a sensible increase in the coefficients with respect to M- and ML estimation. Moreover, it is important to remark that these estimates are somewhat less robust to model specification, with two rather distant values in the models with and without AC-RI interactions (1.267 and 1.161, respectively). As for RI effects on GAS intake, the comparison with the other two approaches yields smaller coefficients, with no significant differences between any pair of levels.

5.2.1 | Causal mediation effects

Since consistency, composition and positivity are technical statements not strictly related to the application context, and the plausibility of assumptions (i)–(iv) has already been discussed in Section 5.2, causal mediation analysis through natural effect decomposition can also be performed. As stated in Section 2.1, in general log odds-ratio NDEs and NIEs can be parametrically estimated with the formulas reported in Section 1 of the Supplementary material. In particular, when the outcome is rare conditionally on the levels of its explanatory variables and the mediator does not interact with the exposure, it can be shown that the nonlinear component of the NDE estimator tends to vanish (see Equation 3 of the Supplementary material). Since this is the case in our applied context, it follows that estimated exposure effects in the outcome model (i.e., RI effects in Table 2) are good approximations of log odds-ratio NDEs. In addition, as there are no mediator–covariate interactions in the outcome model and no covariates in the mediator model, log odds-ratio NIEs depend very little on the covariate patterns (with differences at the sixth digit), even in the setting with interactions (again, refer to Supplementary material Section 1 for details).

With regard to M- and ML estimation, the NIEs related to the RI = 2 versus RI = 1 contrast are always very close to 0 and not significant. This is due to the almost null difference between the effects of these two exposure levels in the mediator model (see the corresponding regression coefficients in Table 1). Conversely, when the reference level of RI is 0, NIEs lie around 0.068, with Delta-method (Oehlert, 1992) standard errors close to 0.042 (p-value 0.105). An exception is represented by the RI = 2 versus RI = 0 contrast in the setting with interactions, where the NIE has the same magnitude but the standard error is 0.037 (p-value 0.066). As for weighting, the log odds-ratio NDE and NIE values are slightly different, reflecting the fact that the estimated regression coefficients differ from those of the other approaches (see the related discussion in Section 5.2). However, the interpretation of results is essentially the same. In Table 3, we report the proportion of RI effect mediated by GAS intake (for the 1 vs. 0 and the 2 vs. 0 contrasts) on the log odds-ratio scale. This proportion is small but nonnegligible, since it ranges from 2.4% to 10.9% across RI contrasts and ACs. Nevertheless, it is worth to remark that the significance of estimated NIEs is weak, and that total causal effects are nonsignificant whenever the corresponding NDEs also are.

TABLE 3 Proportion of reduced immunocompetency (RI) effect mediated by gastric acid suppressor (GAS) intake.

	M-estimation		ML		Weighting	
RI contrast	**1 vs. 0**	**2 vs. 0**	**1 vs. 0**	**2 vs. 0**	**1 vs. 0**	**2 vs. 0**
	Outcome model without interactions					
AGE = All	0.050	0.030	0.050	0.030	0.035	0.027
	Outcome model with interactions					
AGE = 40–65	0.074	0.024	0.072	0.024	0.069	0.025
AGE = 66–75	0.032	0.031	0.033	0.032	0.025	0.032
AGE = 76+	0.108	0.051	0.109	0.053	0.079	0.046

Abbreviation: ML, maximum likelihood.

6 | SIMULATION STUDY

In this section, we present a simulation study conducted to investigate the finite-sample properties of the M- and ML estimators introduced in Section 4. The performance of these estimators is compared to that of the weighting estimator, implemented for both the outcome and mediator model. The target parameters are causal natural direct and indirect effects, that can be computed as detailed in the previous sections. In Section 6.1, the structure of the study is described, while results are reported in Section 6.2.

6.1 | Design

The simulation study mimics the two models of the case study in Section 5 (i.e., with and without RI–AC interactions) by using the same distribution of the covariates and taking population parameter values for (4) and (5) quite close to those estimated from that data set. The only exception concerns the outcome model intercept, which is raised in order to get a slightly higher expected prevalence. In this way, at least 100 cases for each simulated population are obtained. We generate 1000 populations of 3 million individuals. For each generated population, we extract both a CC and an SCC sample, the latter involving stratification on ACs. In detail, we randomly select 100 or 50 cases, and for each case we randomly select five or eight controls from its AC (SCC setting) or from the entire population of controls (CC setting). Notice that for each repetition marginal as well as AC-specific population prevalences can be computed exactly.

For every combination of the above factors (number of cases, number of controls per case, and CC or SCC setting), four scenarios are considered. In all scenarios, RI and AC are simulated randomly with $P(\text{RI}=0)=0.73$, $P(\text{RI}=1)=0.16$, $P(\text{RI}=2)=0.11$, $P(\text{AC}=1)=0.66$, $P(\text{AC}=2)=0.19$, $P(\text{AC}=3)=0.15$. Furthermore, the mediator (GAS intake) is generated by $M \sim \text{Be}(\text{expit}(\boldsymbol{x}_m^\top \boldsymbol{\delta}))$, with $\boldsymbol{x}_m^\top = (1, \mathbb{I}\{\text{RI}=1\}, \mathbb{I}\{\text{RI}=2\})$ and $\boldsymbol{\delta} = (-3.3, 0.8, 0.8)^\top$. On the other hand, the outcome (listeriosis onset) is generated in two different ways, by using different $\boldsymbol{x}_y$ and $\boldsymbol{\beta}$ to generate $Y \sim \text{Be}(\text{expit}(\boldsymbol{x}_y^\top \boldsymbol{\beta}))$. In Scenarios 1 and 2 (corresponding to the no interaction model), $\boldsymbol{x}_{y_{12}}$ is as in the top part of Table 1 and $\boldsymbol{\beta_{12}} = (-8.5, 1.3, 2.2, 1.0, 0.7, 1.0)^\top$. In Scenarios 3 and 4 (corresponding to the RI-AC interaction model), $\boldsymbol{x}_{y_{34}}$ is as in the bottom part of Table 1 and $\boldsymbol{\beta_{34}} = (-9, 0.9, 2.7, 0.9, 1.5, 1.2, -0.6, -0.5, -1.6, 1)^\top$. In order to test the estimators' performance in the presence of model misspecification, in Scenarios 2 and 4 we have added a normally distributed variable with mean 0 and standard deviation 2 in the estimation of the outcome model; this variable was not there in the data-generating process and therefore the logit link is no longer an appropriate link.

Since the conditions mentioned in Section 5.2.1 (no exposure–mediator and mediator–covariate interactions in the outcome model, conditional outcome rareness) still hold in this simulation setting, the benchmark log odds-ratio NDEs are very close to the conditional RI effects that can be reconstructed from the regression coefficients, while log odds-ratio NIEs are almost invariant to AC patterns. Specifically, in Scenarios 1 and 2 all NDEs are very close to 1.3 (for RI = 1 vs. RI = 0) and to 2.2 (for RI = 2 vs. RI = 0), whereas in Scenarios 3 and 4 the reference values for the three ACs are 0.9, 2.1, and 0.4 (for RI = 1 vs. RI = 0) and 2.7, 2.1, and 1.1 (for RI = 2 vs. RI = 0). Moreover, for all scenarios and contrasts log odds-ratio NIEs are very close to 0.063. With this respect, it is worth to clarify that in Scenarios 2 and 4 estimated natural effects are computed by setting the additional normal covariate in the outcome model to 0. However, this choice is essentially irrelevant, since such a covariate does not interact either with RI or AC, so all effects depend very little on its level.

6.2 | Results

Table 4 summarizes the performance metrics of the three estimators of log odds-ratio NDEs and NIEs for the CC setting with eight controls per case. Analogous tables for the CC setting with five controls per case and for the SCC setting (with both five and eight controls per case) are reported in Section 3 of the Supplementary material (Tables 2–4). In all these tables, effects refer to the RI = 2 versus RI = 0 contrast.

For a few repetitions, when using 50 cases, the maximization was unstable, resulting in nonconvergence of the regression models. This was more common for M-estimation and weighting, however for these repetitions the ML estimates were also nearly singular and therefore rendered very unstable estimates. To make comparisons between estimators fair, we removed the repetitions (between 0 and 5 depending on sample size and scenario) for which one or more method was not able to estimate both the regression parameters and the standard errors of the regression parameters (see Table 5 in the Supplementary material for the exact number of removed repetitions).

In all configurations, the NIE is very small compared to the NDE, therefore the behavior of the TE (not shown) is quite close to that of the NDE. There are no dramatic differences between the estimators of the NDE: ML and M-estimation always outperform weighting in Scenarios 1 and 2 (though with different intensities), whereas in Scenarios 3 and 4 there is a less univocal pattern (see results across the four tables). For these estimators, empirical coverage is always close to the nominal level, even in the presence of bias.

The ML and M-estimators of NIEs are often similar and show, for each configuration, lower bias and root mean square error compared to the weighting estimator, especially when the number of cases is lower. Regardless of the estimation method, slight systematic undercoverage occurs, suggesting that 95% confidence intervals should be interpreted with caution. This is unsurprising since NIEs are nonlinear functions of the regression parameters, and therefore their estimates are more prone to deviate from the normal distribution in finite samples. Conversely, NDEs suffer considerably less from this issue. This is due to the absence of exposure–mediator interactions and to the conditional rareness of the outcome, thanks to which the nonlinear component of estimated NDEs tends to be negligible (see Section 5.2.1).

7 | DISCUSSION

We consider SCC designs defined on a certain variable (primary outcome) and we address the problem of modeling another binary variable (secondary outcome) from the resulting data set. In detail, we assume that a logistic regression model for the secondary outcome holds in the target population, and we recover the parameters of such a model via a simple offset adjustment; see Sections 3 and 4. Like in other frameworks, knowledge of the (conditional) prevalence(s) of the primary outcome in the population is required.

Our analytical solution allows one to perform parametric causal mediation analysis from SCC data in settings where both the outcome and the mediator are binary. However, the identification strategy can be exploited also in a nonparametric framework. With reference to parametric estimation, we have shown in Section 4 how the derived offset correction leads to M- and ML estimation of the joint (i.e., outcome and mediator) model parameter vector, provided that the intercept and/or the coefficients of background variables in the outcome model are suitably adjusted. These estimation methods are opposed to weighting, where such adjustments are not needed. It is worth to mention that the general ML framework by Imbens (1992) can also account for the setting (not considered here) where the primary outcome prevalences in the population have to be estimated along with other parameters. However, in that case the procedure described in Section 4.2 would need modifications in order to return suitable ML estimates and standard errors.

Our simulation study, which targets log odds-ratio NDEs and NIEs, shows that M- and ML estimation typically outperform weighting (though with exceptions), even in the presence of model misspecification. In particular, better performances are observed for NIEs, a finding suggesting that the efficiency gains of these two estimation methods extend to secondary outcome models.

The proposed approach is exemplified via the analysis of data gathered within a German study on listeriosis conducted by Preußel et al. (2015). Specifically, we take a novel perspective and focus on evaluating whether the effect of RI on listeriosis development is mediated by the intake of GASs. We try to answer this question also in causal terms by decomposing the total causal effect of RI on listeriosis into the NDE and the NIE, though we acknowledge that such a terminology might be somewhat questionable from a strict causal mediation standpoint. This is because RI and GAS intake do not indeed *cause* listeriosis; the ingestion of food contaminated with *L. monocytogenes* bacteria (predominantly) or other unknown

TABLE 4 Root mean square error (RMSE), Bias, and Empirical coverage of 95% confidence intervals for estimates of log odds-ratio natural direct effects (NDEs) and natural indirect effects (NIEs) (RI = 2 vs. RI = 0 contrast, case–control (CC) setting with eight controls per case). Confidence intervals are constructed assuming normal distribution (which is not always appropriate, see comments in text) and standard errors are computed with sandwich estimators. For Scenarios 1 and 2, we report result for age class (AC) = 3, with those for other ACs being almost identical. The weighting estimator is abbreviated to W, and *n* stands for the number of cases.

			RMSE			Bias			Empirical coverage		
Effect	***n***	**AC**	**M-est**	**ML**	**W**	**M-est**	**ML**	**W**	**M-est**	**ML**	**W**
Scenario 1: No interaction, correct model											
Direct	50	3	0.40	0.40	0.43	0.03	0.03	0.06	0.95	0.95	0.95
	100		0.26	0.26	0.27	0.03	0.03	0.04	0.96	0.96	0.96
Indirect	50		0.06	0.06	0.09	−0.01	0.00	−0.01	0.91	0.91	0.92
	100		0.04	0.04	0.05	−0.01	−0.01	−0.01	0.91	0.90	0.93
Scenario 2: No interaction, misspecified model											
Direct	50	3	0.39	0.39	0.42	0.07	0.07	0.11	0.96	0.96	0.95
	100		0.26	0.26	0.27	0.03	0.03	0.04	0.96	0.96	0.96
Indirect	50		0.06	0.06	0.10	−0.01	−0.01	−0.02	0.91	0.90	0.91
	100		0.04	0.04	0.05	0.00	−0.01	−0.01	0.91	0.91	0.93
Scenario 3: Interaction, correct model											
Direct	50	1	0.58	0.58	0.60	0.07	0.07	0.07	0.97	0.97	0.97
		2	2.02	2.01	1.93	0.24	0.24	0.27	0.95	0.97	0.96
		3	3.14	3.11	2.99	−0.41	−0.40	−0.35	0.94	0.97	0.94
	100	1	0.39	0.39	0.40	0.02	0.01	0.01	0.97	0.96	0.96
		2	0.61	0.61	0.62	0.06	0.07	0.08	0.95	0.95	0.95
		3	0.85	0.85	0.85	−0.04	−0.03	−0.03	0.96	0.96	0.96
Indirect	50	1	0.06	0.06	0.12	−0.01	−0.01	−0.01	0.91	0.90	0.92
		2	0.06	0.06	0.12	−0.01	−0.01	−0.01	0.91	0.90	0.92
		3	0.06	0.06	0.12	−0.01	−0.01	−0.01	0.91	0.90	0.92
	100	1	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94
		2	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94
		3	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94
Scenario 4: Interaction, misspecified model											
Direct	50	1	0.84	0.84	0.83	0.08	0.07	0.08	0.96	0.96	0.96
		2	1.46	1.48	1.43	0.13	0.13	0.17	0.97	0.97	0.96
		3	3.18	3.13	2.97	−0.47	−0.44	−0.40	0.93	0.96	0.93
	100	1	0.39	0.39	0.40	0.02	0.01	0.01	0.97	0.96	0.96
		2	0.61	0.61	0.62	0.06	0.07	0.08	0.95	0.95	0.95
		3	0.85	0.85	0.85	−0.04	−0.03	−0.03	0.96	0.96	0.96
Indirect	50	1	0.06	0.06	0.12	−0.01	−0.01	−0.01	0.93	0.92	0.94
		2	0.06	0.06	0.11	−0.01	−0.01	−0.01	0.93	0.92	0.94
		3	0.06	0.06	0.12	−0.01	−0.01	−0.01	0.93	0.92	0.94
	100	1	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94
		2	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94
		3	0.04	0.04	0.05	−0.01	−0.01	−0.01	0.93	0.93	0.94

Abbreviation: ML, maximum likelihood.

vehicles (to a lesser extent) do. Nevertheless, since infection means are not of primary interest, the disentanglement of the RI effect on listeriosis in causal terms appears sensible, due to the clear additional vulnerability to listeriosis for people with RI as well as for people taking GASs, and to the fact that GAS intake is a direct consequence of RI. Results show that the NIE of RI is almost negligible.

It is important to stress that any conclusion from the reported analyses has to be drawn cautiously, for several reasons. First, as mentioned in Section 5.1 only 99 of 732 patients potentially eligible as cases formed the final data set due to participation refusal, decease, inability to answer/be contacted, and other unknown sources of missingness. Although this fact is taken into account by generating proper (S)CC samples via random selection of controls, it is important to remark that like in many telephone surveys controls were found to systematically differ from the target population with respect to socioeconomic status. While socioeconomic status is essentially accounted for via the (initial) inclusion of educational level as its proxy (Preußel et al., 2015), we cannot exclude that some degree of noninformative missingness (Molenberghs et al., 2014) might affect the results.

Furthermore, though as discussed in Section 5.2 the typical assumptions of causal mediation analysis (i)–(iv) seem to be tenable in our application, it is worth to mention that a number of approaches exist which deal with possible violations. As for unobserved confounding (i)–(iii), which is quite often due to some relevant unmeasured variables, sensitivity analyses (Lindmark et al., 2018) as well as methods embedded in an instrumental variable framework (Didelez, Meng et al., 2010; Mattei & Mealli, 2011) have been introduced which could be possibly adapted to the present setting. Also, the cross-world independence assumption (iv) is typically difficult to interpret (see Section 2.1). With this regard, alternative causal estimands identifiable without making the cross-world independence assumption (named *interventional effects*) were proposed (VanderWeele et al., 2014; Vansteelandt & Daniel, 2017).

It is well-known that in SCC designs the number of strata can in principle increase until the limit situation of containing a single case (sometimes termed *exact matching*). In these settings, traditional estimation might be problematic, with conditional likelihood methods typically preferable (Breslow, 1996; Gail et al., 1981). In particular, links between causal mediation on the hazard ratio scale for time-to-event outcomes and such a conditional likelihood framework have been recently explored in a setting assuming events are rare (Kim et al., 2020). Extensions beyond this assumption, possibly building on the analytic derivations proposed here, might be of interest. Further developments may involve models with latent mediators, which have been recently considered in the literature, too (Albert et al., 2016).

ACKNOWLEDGMENTS

We thank three anonymous reviewers for their constructive comments. We also acknowledge the financial support from Swedish Research Council for health working life and welfare [2019-01064].

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Marco Doretti https://orcid.org/0000-0002-1050-4742

REFERENCES

Ahrens, D., Behrens, G., Himmel, W., Kochen, M., & Chenot, J.-F. (2012). Appropriateness of proton pump inhibitor recommendations at hospital discharge and continuation in primary care. *International Journal of Clinical Practice*, *66*(8), 767–773.

Albert, J. M., Geng, C., & Nelson, S. (2016). Causal mediation analysis with a latent mediator. *Biometrical Journal*, *58*(3), 535–548.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, *59*(1), 19–35.

Andrews, R. M., & Didelez, V. (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, *32*(2), 209–219.

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. In: *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 357–363).

Bavishi, C., & Dupont, H. (2011). Systematic review: The use of proton pump inhibitors and increased susceptibility to enteric infection. *Alimentary Pharmacology & Therapeutics*, *34*(11-12), 1269–1281.

Breslow, N. E. (1996). Statistics in rpidemiology: The case-control study. *Journal of the American Statistical Association*, *91*(433), 14–28.

Breslow, N. E., Zhao, L. P., Fears, T. R., & Brown, C. C. (1988). Logistic regression for stratified case-control studies. *Biometrics*, *44*(3), 891–899.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31.

Didelez, V., Kreiner, S., & Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, *25*(3), 368–387.

Didelez, V., Meng, S., & Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, *25*(1), 22–40.

Doretti, M., Raggi, M., & Stanghellini, E. (2022). Exact parametric causal mediation analysis for a binary outcome with a binary mediator. *Statistical Methods & Applications*, *31*(1), 87–108.

Fears, T. R., & Brown, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, *42*(4), 955–960.

Gail, M. H., Lubin, J. H., & Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, *68*(3), 703–707.

Geneletti, S., Richardson, S., & Best, N. (2009). Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*, *10*(1), 17–31.

Goulet, V., Hebert, M., Hedberg, C., Laurent, E., Vaillant, V., De Valk, H., & Desenclos, J.-C. (2012). Incidence of listeriosis and related mortality among groups at risk of acquiring listeriosis. *Clinical Infectious Diseases*, *54*(5), 652–660.

Henningsen, A., & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, *26*(3), 443–458.

Ho, J. L., Shands, K. N., Friedland, G., Eckind, P., & Fraser, D. W. (1986). An outbreak of type 4b Listeria monocytogenes infection involving patients from eight Boston hospitals. *Archives of Internal Medicine*, *146*(3), 520–524.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*(1), 73–101.

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

Imbens, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, *60*(5), 1187–1214.

Kartsonaki, C., & Cox, D. (2023). Regression reconstruction from a retrospective sample. *Econometrics and Statistics*, *25*, 87–92.

Kim, Y. M., Cologne, J. B., Jang, E., Lange, T., Tatsukawa, Y., Ohishi, W., Utada, M., & Cullings, H. M. (2020). Causal mediation analysis in nested case-control studies using conditional logistic regression. *Biometrical Journal*, *62*(8), 1939–1959.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Lindmark, A., de Luna, X., & Eriksson, M. (2018). Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. *Statistics in Medicine*, *37*(10), 1744–1762.

Manski, C. F., & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, *45*(8), 1977–1988.

Manski, C. F., & McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In C. F. Manski, & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 2–50). MIT Press.

Mattei, A., & Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 729–752.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Mook, P., Jenkins, J., O'Brien, S., & Gillespie, I. (2013). Existing medications among non-pregnancy-related listeriosis patients in England, 2007–2009. *Epidemiology & Infection*, *141*(1), 36–44.

Nguyen, T. Q., Dafoe, A., & Ogburn, E. L. (2019). The magnitude and direction of collider bias for binary variables. *Epidemiologic Methods*, *8*(1), 20180011.

Oehlert, G. W. (1992). A note on the delta method. *American Statistician*, *46*(1), 27–29.

Pearl, J. (2001). Direct and indirect effects. In *UAI' 01: Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). Morgan Kaufmann Publishers Inc.

Pearl, J. (2010). *The mediation formula: A guide to the assessment of causal pathways in non-linear models* (Technical Report R-363). University of California.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403–411.

Preußel, K., Milde-Busch, A., Schmich, P., Wetzstein, M., Stark, K., & Werber, D. (2015). Risk factors for sporadic non-pregnancy associated listeriosis in Germany—Immunocompromised patients and frequently consumed ready-to-eat products. *PLoS One*, *10*(11), e0142986.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, *10*(2), 462–474.

Richardson, D. B., Rzehak, P., Klenk, J., & Weiland, S. K. (2007). Analyses of case–control data for additional outcomes. *Epidemiology*, *18*(4), 441–445.

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

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, *54*(2), 221–226.
Satten, G. A., Curtis, S. W., Solis-Lemus, C., Leslie, E. J., & Epstein, M. P. (2022). Efficient estimation of indirect effects in case-control studies using a unified likelihood framework. *Statistics in Medicine*, *41*(15), 2879–2893.
Stanghellini, E., & Doretti, M. (2019). On marginal and conditional parameters in logistic regression models. *Biometrika*, *106*(3), 732–739.
Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. *American Statistician*, *56*(1), 29–38.
Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine*, *32*(26), 4567–4580.
Tchetgen Tchetgen, E. J., & VanderWeele, T. J. (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, *25*(2), 282–291.
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*, *18*(2), 137–150.
van der Laan, M. J. (2008). Estimation based on case-control designs with known prevalence probability. *International Journal of Biostatistics*, *4*(1), Article No. 17.
VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883.
VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.
VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2016). Mediation analysis with matched case-control study designs. *American Journal of Epidemiology*, *183*(9), 869–870.
VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, intervention and composition. *Statistics and Its Interface*, *2*(4), 457–468.
VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*(12), 1339–1348.
VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, *25*(2), 300–306.
Vansteelandt, S., & Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, *28*(2), 258–265.
Wang, J., & Shete, S. (2011). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology*, *35*(3), 190–200.
Wang, J. J. J., Bartlett, M., & Ryan, L. (2017). Non-ignorable missingness in logistic regression. *Statistics in Medicine*, *36*(19), 3005–3021.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.
Xie, Y., & Manski, C. F. (1989). The logit model and response-based samples. *Sociological Methods & Research*, *17*(3), 283–302.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Doretti, M., Genbäck, M., & Stanghellini, E. (2024). Mediation analysis with case–control sampling: Identification and estimation in the presence of a binary mediator. *Biometrical Journal*, *66,* 2300089. https://doi.org/10.1002/bimj.202300089

APPENDIX A: DETAILS OF PARAMETRIC AND NONPARAMETRIC IDENTIFICATION

By standard probability results it is possible to express the left-hand side of (7) as

$$\begin{aligned}\log\frac{P(Y=1\mid a,b,z,m,W=1)}{P(Y=0\mid a,b,z,m,W=1)} &= \log\frac{P(W=1\mid a,b,z,m,Y=1)P(Y=1\mid a,b,z,m)}{P(W=1\mid a,b,z,m,Y=0)P(Y=0\mid a,b,z,m)}\\ &= \operatorname{logit}\{P(Y=1\mid a,b,z,m)\}+\log\frac{P(W=1\mid Y=1,B=b)}{P(W=1\mid Y=0,B=b)}\\ &= \boldsymbol{x}_y^\top\boldsymbol{\beta}+\log(k_b).\end{aligned}\tag{A1}$$

The second equality follows from $W \perp\!\!\!\perp (A,Z,M)\mid(Y,B)$, which holds true in the directed acyclic graph (DAG) in Figure 1(a). Clearly, the compact form in the right-hand side of (7) can be used in place of the last term in the above equation chain.

The same approach allows to express the left-hand side of (9) as

$$\begin{aligned}\log\frac{P(M=1\mid a,b,z,y)}{P(M=0\mid a,b,z,y)} &= \log\frac{P(Y=y\mid a,b,z,M=1)}{P(Y=y\mid a,b,z,M=0)} + \log\frac{P(M=1\mid a,b,z)}{P(M=0\mid a,b,z)}\\ &= \log\frac{P(Y=y\mid a,b,z,M=1)}{P(Y=y\mid a,b,z,M=0)} + \log\frac{P(M=1\mid a,b,z)}{P(M=0\mid a,b,z)}\\ &= o(y,\boldsymbol{x}_y;\boldsymbol{\beta}) + \boldsymbol{x}_m^\top\boldsymbol{\delta}.\end{aligned} \tag{A2}$$

To obtain the parametric expression of $o(y,\boldsymbol{x}_y;\boldsymbol{\beta})$, the partition $\boldsymbol{x}_y = (\boldsymbol{x}_{y_0}^\top, \boldsymbol{x}_{y_1}^\top)^\top$ is introduced, where $\boldsymbol{x}_{y_0}$ ($\boldsymbol{x}_{y_1}$) denotes the subvector of covariate values not involving (involving) M. The coefficient vector partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top)^\top$ is defined accordingly. We assume our model to be hierarchical, and therefore in the most general setting $\boldsymbol{x}_{y_1} = m\cdot\boldsymbol{x}_{y_0}$. However, in many applications $\boldsymbol{\beta}_1$ and $\boldsymbol{x}_{y_1}$ are likely to be smaller dimensional than $\boldsymbol{\beta}_0$ and $\boldsymbol{x}_{y_0}$, with some elements of the latter vectors not having a counterpart in the former. In these cases, the additional vector $\tilde{\boldsymbol{\beta}}_1$ has to be defined by suitably expanding $\boldsymbol{\beta}_1$ with zeros, so that the sum of conformable vectors $\boldsymbol{\beta}_+ = \boldsymbol{\beta}_0 + \tilde{\boldsymbol{\beta}}_1$ can be computed. It then follows that

$$\begin{aligned}\text{logit}\{P(Y=1\mid a,b,z,M=0)\} &= \boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_0\\ \text{logit}\{P(Y=1\mid a,b,z,M=1)\} &= \boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_+.\end{aligned}$$

Then,

$$\begin{aligned}o(y,\boldsymbol{x}_y;\boldsymbol{\beta}) = \log\frac{P(Y=y\mid a,b,z,M=1)}{P(Y=y\mid a,b,z,M=0)} &= \log\frac{\exp\{y\cdot\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_+\}}{1+\exp\{\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_+\}} - \log\frac{\exp\{y\cdot\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_0\}}{1+\exp\{\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_0\}}\\ &= y(\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_+ - \boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_0) - \log\frac{1+\exp\{\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_+\}}{1+\exp\{\boldsymbol{x}_{y_0}^\top\boldsymbol{\beta}_0\}}.\end{aligned} \tag{A3}$$

To achieve nonparametric identification of $P(Y=y\mid a,b,z,m)$ and of $P(M=m\mid a^\star,b,z)$, the following algorithm can be implemented:

1. Consider the observable probabilities $P(Y=1\mid a,b,z,m,W=1)$ and $P(Y=1\mid a^\star,b,z,W=1)$;
2. Take the odds transforms of the quantities in 1;
3. Use knowledge of k_b to identify the odds transforms of $P(Y=1\mid a,b,z,m)$ and $P(Y=1\mid a^\star,b,z)$ via

$$\frac{P(Y=1\mid a,b,z,m)}{P(Y=0\mid a,b,z,m)} = \frac{P(Y=1\mid a,b,z,m,W=1)}{P(Y=0\mid a,b,z,m,W=1)}\times\frac{1}{k_b}$$

and

$$\frac{P(Y=1\mid a^\star,b,z)}{P(Y=0\mid a^\star,b,z)} = \frac{P(Y=1\mid a^\star,b,z,W=1)}{P(Y=0\mid a^\star,b,z,W=1)}\times\frac{1}{k_b}$$

(see Equation A1 for reference);
4. Revert to the probability scale to identify $P(Y=1\mid a,b,z,m)$ and $P(Y=1\mid a^\star,b,z)$;
5. For $y,m\in\{0,1\}$, consider $P(M=m\mid a^\star,b,z,y,W=1) = P(M=m\mid a^\star,b,z,y)$;
6. Use the quantities in 4 and 5 to identify $P(M=1\mid a^\star,b,z)$ by using

$$P(M=m\mid a^\star,b,z) = \sum_y P(M=m\mid a^\star,b,z,y)P(Y=y\mid a^\star,b,z).$$

Importantly, the identities in point 3 hold true only in consequence of Bayes' theorem paired with the conditional independence statement $W \perp\!\!\!\perp (A,Z,M)\mid(Y,B)$, and they do not depend on any parametric model. Nonparametric estimation of the above probabilities can be performed with sample relative frequencies or, in the case of continuous A

and/or continuous or high-dimensional Z, with any other suitable nonparametric method. Notice that this strategy can also be adapted to address continuous mediators.

APPENDIX B: DETAILS OF M-ESTIMATION

In order to obtain the expressions of the observed score function vectors $\boldsymbol{s}_y(\cdot)$ and $\boldsymbol{s}_m(\cdot)$, it is worth to rely on matrix notation. Specifically, we index sample units by $i = 1, \dots, n$ and introduce the sample vectors $\boldsymbol{m} = (m_1, \dots, m_n)^\top$ and $\boldsymbol{y} = (y_1, \dots, y_n)^\top$, as well as the sample design matrices

$$\boldsymbol{X}_m = \begin{pmatrix} \boldsymbol{x}_{m,1}^\top \\ \vdots \\ \boldsymbol{x}_{m,n}^\top \end{pmatrix}, \qquad \boldsymbol{X}_y = \begin{pmatrix} \boldsymbol{X}_{y_0} & \boldsymbol{X}_{y_1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{y_0,1}^\top & \boldsymbol{x}_{y_1,1}^\top \\ \vdots & \vdots \\ \boldsymbol{x}_{y_0,n}^\top & \boldsymbol{x}_{y_1,n}^\top \end{pmatrix}.$$

We also denote by d_δ, d_β, d_{β_0}, and d_{β_1} the number of columns of $\boldsymbol{X}_m$, $\boldsymbol{X}_y$, $\boldsymbol{X}_{y_0}$, and $\boldsymbol{X}_{y_1}$, respectively.

Letting $\text{expit}(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$, for a given value $\theta = (\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top)^\top$ we have

$$\begin{aligned} \boldsymbol{s}_y(\theta) &= \boldsymbol{X}_y^\top(\boldsymbol{y} - \boldsymbol{p}_{y^\star}) & \boldsymbol{p}_{y^\star} &= \text{expit}(\boldsymbol{\eta}_{y^\star}) & \boldsymbol{\eta}_{y^\star} &= \boldsymbol{X}_y\boldsymbol{\beta}^\star \\ \boldsymbol{s}_m(\theta) &= \boldsymbol{X}_m^\top(\boldsymbol{m} - \boldsymbol{p}_{my}) & \boldsymbol{p}_{my} &= \text{expit}(\boldsymbol{\eta}_{my}) & \boldsymbol{\eta}_{my} &= \boldsymbol{X}_m\boldsymbol{\delta} + \boldsymbol{o}, \end{aligned}$$

where $\boldsymbol{\beta}^\star$ is like in Section 3 and $\boldsymbol{o}$ is a column vector collecting the offset terms $o(y_i, \boldsymbol{x}_{y_0,i}; \boldsymbol{\beta})$ of every sample unit $i = 1, \dots, n$. Notice that the vector $\boldsymbol{p}_{my} = (p_{my,1}, \dots, p_{my,n})^\top$ contains the $p_{my,i}$ probabilities ($i = 1, \dots, n$) already introduced in Section 4.2, while $\boldsymbol{p}_{y^\star} = (p_{y^\star,1}, \dots, p_{y^\star,n})^\top$ differs from $\boldsymbol{p}_{y^\star_{(m)}} = (p_{y^\star_{(m)},1}, \dots, p_{y^\star_{(m)},n})$, which is the vector collecting the $p_{y^\star_{(m)},i}$ probabilities (also used in Section 4.2). A compact form for $\boldsymbol{\psi}(\theta) = (\boldsymbol{s}_y(\theta)^\top, \boldsymbol{s}_m(\theta)^\top)^\top$ is given by

$$\boldsymbol{\psi}(\theta) = \boldsymbol{X}_j^\top \boldsymbol{f},$$

where

$$\boldsymbol{X}_j = \begin{pmatrix} \boldsymbol{X}_y & \boldsymbol{0}_{n\times d_\delta} \\ \boldsymbol{0}_{n\times d_\beta} & \boldsymbol{X}_m \end{pmatrix} \qquad \boldsymbol{f} = \begin{pmatrix} \boldsymbol{y} - \boldsymbol{p}_{y^\star} \\ \boldsymbol{m} - \boldsymbol{p}_{my} \end{pmatrix}.$$

The formula for the $A(\boldsymbol{y}, \theta)$ matrix is

$$A(\boldsymbol{y}, \theta) = \frac{1}{n}\begin{pmatrix} \boldsymbol{H}_{y\beta} & \boldsymbol{H}_{y\delta} \\ \boldsymbol{H}_{m\beta} & \boldsymbol{H}_{m\delta} \end{pmatrix},$$

where each generic submatrix is defined as $\boldsymbol{H}_{q\alpha} = \partial \boldsymbol{s}_q / \partial \boldsymbol{\alpha}^\top$ and

$$\begin{aligned} \boldsymbol{H}_{y\beta} &= -\boldsymbol{X}_y^\top \text{diag}\{\boldsymbol{p}_{y^\star} \cdot (1 - \boldsymbol{p}_{y^\star})\}\boldsymbol{X}_y & \boldsymbol{H}_{y\delta} &= \boldsymbol{0}_{d_\beta \times d_\delta} \\ \boldsymbol{H}_{m\beta} &= -\boldsymbol{X}_m^\top \text{diag}\{\boldsymbol{p}_{my} \cdot (1 - \boldsymbol{p}_{my})\}\boldsymbol{D} & \boldsymbol{H}_{m\delta} &= -\boldsymbol{X}_m^\top \text{diag}\{\boldsymbol{p}_{my} \cdot (1 - \boldsymbol{p}_{my})\}\boldsymbol{X}_m. \end{aligned}$$

In the above, the $\boldsymbol{D}$ matrix is given by

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{0}_{n\times d_{\beta_0}} & \bar{\boldsymbol{X}}_{y_0} \cdot \boldsymbol{y} \end{pmatrix} + \begin{pmatrix} \boldsymbol{X}_{y_0} \cdot \boldsymbol{v}_0 & \bar{\boldsymbol{X}}_{y_0} \cdot \boldsymbol{v}_1 \end{pmatrix},$$

where $\bar{\boldsymbol{X}}_{y_0}$ is the matrix obtained by extracting the columns of $\boldsymbol{X}_{y_0}$ corresponding to the nonnull elements of $\tilde{\boldsymbol{\beta}}_1$ while $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ are two column vectors of length n, whose ith element is given by

$$v_{0,i} = \frac{\exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_0) - \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_+)}{\{1 + \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_0)\}\{1 + \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_+)\}} \qquad v_{1,i} = -\frac{\exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_+)}{1 + \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_+)}.$$

Notice that in the formulas above · denotes element-wise product, be it between column vectors or between a matrix and a column vector. In the latter case, the result is a conformable matrix where every column is element-wise multiplied by the vector.

Finally, $B(\boldsymbol{y}, \theta)$ is given by

$$B(\boldsymbol{y}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}_i(\theta) \boldsymbol{\psi}_i^\top(\theta),$$

with $\boldsymbol{\psi}_i(\theta)$ being the contribution of unit i to the observed score vector.

APPENDIX C: PARAMETRIC EXPRESSION OF (14)

The parametric expression of $g(\boldsymbol{x}_{y_0}, \boldsymbol{x}_m; \boldsymbol{\beta}, \boldsymbol{\delta})$ can be derived from (A2) and (A3). After some simplifications, it follows that

$$g(\boldsymbol{x}_{y_0}, \boldsymbol{x}_m; \boldsymbol{\beta}, \boldsymbol{\delta}) = \log \frac{\exp(\boldsymbol{x}_{y_0}^\top \tilde{\boldsymbol{\beta}}_1) \exp(\boldsymbol{x}_m^\top \boldsymbol{\delta}) \{1 + \exp(\boldsymbol{x}_{y_0}^\top \boldsymbol{\beta}_0)\} + \{1 + \exp(\boldsymbol{x}_{y_0}^\top \boldsymbol{\beta}_+)\}}{\exp(\boldsymbol{x}_m^\top \boldsymbol{\delta}) \{1 + \exp(\boldsymbol{x}_y^\top \boldsymbol{\beta}_0)\} + \{1 + \exp(\boldsymbol{x}_y^\top \boldsymbol{\beta}_+)\}}, \tag{C1}$$

where the elements in the right-hand side are defined as in Appendix A. We then have

$$\text{logit}\{P(Y = 1 \mid a, z, b, W = 1)\} = \boldsymbol{x}_{y_0}^\top \boldsymbol{\beta}_0^\star + g(\boldsymbol{x}_{y_0}, \boldsymbol{x}_m; \boldsymbol{\beta}, \boldsymbol{\delta}),$$

with $\boldsymbol{\beta}_0^\star$ being the modification of $\boldsymbol{\beta}_0$ accounting for the correction for β_{INT} and $\boldsymbol{\beta}_{\text{B}}$.

APPENDIX D: GRADIENT FUNCTION FOR MAXIMUM LIKELIHOOD (ML) ESTIMATION

The log-likelihood gradient $\boldsymbol{s}(\theta)$ can be conveniently expressed in matrix form. To this end, we maintain notation as in the body of the paper and in previous appendices, and we introduce the matrices

$$\boldsymbol{A}_g = \begin{pmatrix} \boldsymbol{A}_{gy_0} \\ \boldsymbol{A}_{gy_1} \\ \boldsymbol{A}_{gm} \end{pmatrix} \qquad \boldsymbol{B}_g = \begin{pmatrix} \boldsymbol{B}_{gy_0} \\ \boldsymbol{B}_{gy_1} \\ \boldsymbol{B}_{gm} \end{pmatrix} \qquad \boldsymbol{A}_m = \begin{pmatrix} \boldsymbol{A}_{my_0} \\ \boldsymbol{A}_{my_1} \\ \boldsymbol{A}_{mm} \end{pmatrix} \qquad \boldsymbol{B}_m = \begin{pmatrix} \boldsymbol{B}_{my_0} \\ \boldsymbol{B}_{my_1} \\ \boldsymbol{B}_{mm} \end{pmatrix},$$

where

$$\begin{aligned}
&\boldsymbol{A}_{gy_0} = \mathbf{1}_{d_{\beta_0} \times n} && \boldsymbol{B}_{gy_0} = \boldsymbol{G}_{\beta_{\text{INT}}} \otimes \mathbf{1}_{d_{\beta_0}} && \boldsymbol{A}_{my_0} = \mathbf{0}_{d_{\beta_0} \times n} && \boldsymbol{B}_{my_0} = \boldsymbol{v}_0^\top \otimes \mathbf{1}_{d_{\beta_0}} \\
&\boldsymbol{A}_{gy_1} = \mathbf{0}_{d_{\beta_1} \times n} && \boldsymbol{B}_{gy_1} = \boldsymbol{G}_{\beta_m} \otimes \mathbf{1}_{d_{\beta_1}} && \boldsymbol{A}_{my_1} = \boldsymbol{y}^\top \otimes \mathbf{1}_{d_{\beta_1}} && \boldsymbol{B}_{my_1} = \boldsymbol{v}_1^\top \otimes \mathbf{1}_{d_{\beta_1}} \\
&\boldsymbol{A}_{gm} = \mathbf{0}_{d_\delta \times n} && \boldsymbol{B}_{gm} = \boldsymbol{G}_{\delta_{\text{INT}}} \otimes \mathbf{1}_{d_\delta} && \boldsymbol{A}_{mm} = \mathbf{1}_{d_\delta \times n} && \boldsymbol{B}_{mm} = \mathbf{0}_{d_\delta \times n}.
\end{aligned}$$

In the above, $\otimes$ denotes Kronecker product, while

$$\begin{aligned}
\boldsymbol{G}_{\beta_{\text{INT}}} &= \left(\frac{\partial}{\partial \beta_{\text{INT}}} g(\boldsymbol{x}_{y_0,1}, \boldsymbol{x}_{m,1}; \boldsymbol{\beta}, \boldsymbol{\delta}) \quad \cdots \quad \frac{\partial}{\partial \beta_{\text{INT}}} g(\boldsymbol{x}_{y_0,n}, \boldsymbol{x}_{m,n}; \boldsymbol{\beta}, \boldsymbol{\delta}) \right) \\
\boldsymbol{G}_{\beta_m} &= \left(\frac{\partial}{\partial \beta_m} g(\boldsymbol{x}_{y_0,1}, \boldsymbol{x}_{m,1}; \boldsymbol{\beta}, \boldsymbol{\delta}) \quad \cdots \quad \frac{\partial}{\partial \beta_m} g(\boldsymbol{x}_{y_0,n}, \boldsymbol{x}_{m,n}; \boldsymbol{\beta}, \boldsymbol{\delta}) \right) \\
\boldsymbol{G}_{\delta_{\text{INT}}} &= \left(\frac{\partial}{\partial \delta_{\text{INT}}} g(\boldsymbol{x}_{y_0,1}, \boldsymbol{x}_{m,1}; \boldsymbol{\beta}, \boldsymbol{\delta}) \quad \cdots \quad \frac{\partial}{\partial \delta_{\text{INT}}} g(\boldsymbol{x}_{y_0,n}, \boldsymbol{x}_{m,n}; \boldsymbol{\beta}, \boldsymbol{\delta}) \right)
\end{aligned}$$

are three row vectors collecting the derivatives of the $g(\boldsymbol{x}_{y_0,i}, \boldsymbol{x}_{m,i}; \boldsymbol{\beta}, \boldsymbol{\delta})$ terms $(i = 1, \ldots, n)$ with respect to β_{INT}, β_m, and δ_{INT}, where the latter two are the coefficient of M in the outcome model and the intercept of the mediation model, respectively. These derivatives can be computed by noticing that the argument of the logarithm can be written as $(q_1 q_2 q_3 + q_4)/(q_2 q_3 + q_4)$, with $q_1 = \exp(\boldsymbol{x}_{y_0,i}^\top \tilde{\boldsymbol{\beta}}_1)$, $q_2 = \exp(\boldsymbol{x}_{m,i}^\top \boldsymbol{\delta})$, $q_3 = 1 + \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_0)$, and $q_4 = 1 + \exp(\boldsymbol{x}_{y_0,i}^\top \boldsymbol{\beta}_+)$.

Consequently, we have

$$\frac{\partial g}{\partial \beta_{\text{INT}}} = \frac{\{q_1 q_2 (q_3 - 1) + q_4 - 1\}(q_2 q_3 + q_4) - (q_1 q_2 q_3 + q_4)\{q_2 (q_3 - 1) + q_4 - 1\}}{\exp(g)(q_2 q_3 + q_4)^2}$$

$$\frac{\partial g}{\partial \beta_m} = \frac{(q_1 q_2 q_3 + q_4 - 1)(q_2 q_3 + q_4) - (q_1 q_2 q_3 + q_4)(q_4 - 1)}{\exp(g)(q_2 q_3 + q_4)^2}$$

$$\frac{\partial g}{\partial \delta_{\text{INT}}} = \frac{(q_1 q_2 q_3)(q_2 q_3 + q_4) - (q_1 q_2 q_3 + q_4)(q_2 q_3)}{\exp(g)(q_2 q_3 + q_4)^2}.$$

Finally, letting

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{X}_{y_0} & \bar{\boldsymbol{X}}_{y_0} & \boldsymbol{X}_m \end{pmatrix}^\top,$$

the log-likelihood gradient is given by

$$s(\theta) = \{(\boldsymbol{A}_g + \boldsymbol{B}_g) \cdot \boldsymbol{C}\}\tilde{\boldsymbol{y}} + \{(\boldsymbol{A}_m + \boldsymbol{B}_m) \cdot \boldsymbol{C}\}\tilde{\boldsymbol{m}}, \tag{D1}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{p}_{y^\star_{(m)}}$, $\tilde{\boldsymbol{m}} = \boldsymbol{m} - \boldsymbol{p}_{my}$, and $\boldsymbol{p}_{my}$ and $\boldsymbol{p}_{y^\star_{(m)}}$ are like in Appendix B.