Daniel Nevo*, Xiaomei Liao and Donna Spiegelman

Estimation and Inference for the Mediation Proportion

https://doi.org/10.1515/ijb-2017-0006

Abstract: In epidemiology, public health and social science, mediation analysis is often undertaken to investigate the extent to which the effect of a risk factor on an outcome of interest is mediated by other covariates. A pivotal quantity of interest in such an analysis is the mediation proportion. A common method for estimating it, termed the "difference method", compares estimates from models with and without the hypothesized mediator. However, rigorous methodology for estimation and statistical inference for this quantity has not previously been available. We formulated the problem for the Cox model and generalized linear models, and utilize a data duplication algorithm together with a generalized estimation equations approach for estimating the mediation proportion and its variance. We further considered the assumption that the same link function hold for the marginal and conditional models, a property which we term "g-linkability". We show that our approach is valid whenever g-linkability holds, exactly or approximately, and present results from an extensive simulation study to explore finite sample properties. The methodology is illustrated by an analysis of pre-menopausal breast cancer incidence in the Nurses' Health Study. User-friendly publicly available software implementing those methods can be downloaded from the last author's website (SAS) or from CRAN (R).

Keywords: mediation proportion, mediation analysis, the difference method, proportion of treatment effect

1 Introduction

In many public health, biological, and biomedical systems, the mechanism that explains how an intervention or exposure affects the outcome of interest is unknown, even after a causal association between the exposure and the outcome is established. It is sometimes hypothesized that there exists a *mediator* that connects the exposure and the outcome, sitting on the causal pathway between the exposure and the outcome. In observational studies, identifying a plausible ideally pre-specified, mediator can strengthen the casual inference of the findings. For example, in an evaluation of the effectiveness of the ongoing, trillion dollar President's Emergency Plan for AIDS Relief (PEPFAR) in reducing HIV incidence and prevention in sub-Saharan Africa, it would strengthen the evidence of a causal inference if it could be shown that a substantial proportion of the reduction in disease incidence in time was mediated by increased programmatic coverage in the region, thus diminishing exogenous time trends as the best explanation for any observed decline.

Several methods have been proposed to assess whether mediation exists and to quantify its magnitude [1–4]. [5] described a sequence of hypothesis tests to assess the evidence in the data for mediation by a specific covariate. They assumed a linear model for the relationship between the outcome and the exposure, both marginally and conditionally on the mediator. They also assumed a linear model for the relationship between the exposure and the mediator. Within the counterfactual framework, the building blocks of mediation analysis are the natural direct effect (NDE), defined as the effect on the outcome when increasing the value of the exposure in one unit while holding the mediator at a fixed level, and the natural indirect effect (NIE), which is the effect on the outcome when the exposure is held fixed but the mediator value is changed as it would

***Corresponding author: Daniel Nevo,** Departments of Biostatistics and Epidemiology, Harvard University T H Chan School of Public Health, Boston, Massachusetts 02115, USA, E-mail: danielnevo@gmail.com
Xiaomei Liao, Departments of Biostatistics and Epidemiology, Harvard University T H Chan School of Public Health, Boston, Massachusetts 02115, USA; Currently employed at AbbVie Inc. North Chicago 60064, Illinois, USA, E-mail: merryliao@gmail.com
Donna Spiegelman, Departments of Biostatistics, Epidemiology, Nutrition and Global Health, Harvard University T H Chan School of Public Health, Boston, Massachusetts 02115, USA, E-mail: stdls@hsph.harvard.edu

have been changed if the exposure value were increased by one unit [6–8]. The sum of the NDE and NIE is the total effect (TE). Under this framework, estimation methods for the TE, NDE and NIE were developed for various statistical models. Other examples include logistic regression [9], zero-inflated regression models [10] and high-dimensional mediators in linear regression with normal errors [11].

One way to estimate mediation is through the *product method* [5]. Another widely used method for assessing mediation is the *difference method* [5, 12, 13]. It quantifies the difference in estimates obtained from separate exposure-outcome relationship models, with and without the mediator. The mediation proportion, defined as the change in the effect of the exposure due to mediation by the mediator relative to the total effect, is a main parameter of interest when performing mediation analysis. An analogous measure in surrogacy analysis, termed *proportion of treatment effect* (PTE), aids researchers in deciding whether an intermediate marker can be used as a surrogate for a final outcome of interest. Quantifying PTE entails statistical questions relevant to those that arise in studying the mediation proportion. When the intermediate and the final outcomes are both binary, confidence intervals for the PTE were developed [14]. A time-to-event final outcome with surrogate biomarkers was also considered by [15], who used a data duplication algorithm in order to estimate the covariance between estimators obtained by separate models. The PTE measure in surrogacy research is still actively used and researched [e.g., 16].

Methods for variance estimation, statistical testing and confidence interval construction of mediation parameters have been suggested by past authors. For the difference method, [14] suggested to use results from the linear model in binary outcome setup to approximate the covariance between the two estimators. Other approximation were described and compared in [4]. For the product method, variance can calculated either using delta method [17] or Goodmans exact variance-of-product formula [18]. However, since the finite samples behavior of the product method estimator was found to be nonnormal, the bootstrap was generally recommended [2, 19], at least for medium or small sample size. For the simulation-based mediation approach proposed by [20], a quasi-Bayesian Monte Carlo method or the bootstrap can be used [20, 21].

While the mediation proportion is a popular measure in mediation analysis [e.g., 22–25], statistical inference for this parameter is not sufficiently developed. The NIE and NDE are well-defined concepts, however, in practice, researchers are often primarily interested in the mediation proportion, as exemplified by the aforementioned papers. In this paper, we provide a framework for mediation analysis in generalized linear models (GLMs). We combine a generalized estimation equations (GEE) approach together with a data duplication algorithm to formulate valid statistical inference under minimal assumptions on the marginal and conditional distribution of the outcome. We discuss situations in which these assumptions should hold, and assess robustness to departures from these assumptions in extensive simulation studies. This paper further provides methods for statistical inference in mediation analysis using the difference method, including studying confidence intervals for the mediation proportion and hypothesis tests. Our investigation of these aspects is expanded beyond GLMs to inference about the mediation proportion for Cox model.

The reminder of this paper is organized as follows. In Section 2, we formulate the models needed for the estimation of the mediation proportion in GLMs. In Section 3, we consider the g-linkability property for common link functions. In Section 4, we present methods for inference for this parameter using the multivariate delta method and a data duplication method that enables consistent variance estimation. In Section 5, we present results from a simulation study. In Section 6, we illustrate the use of the methodology developed in studying mediation of the effect of risk factors for pre-menopausal breast cancer incidence by mammographic density in the Nurses' Health Studies NHSI and NHSII. In Section 7, we discuss results and related issues. We describe the software we have made publicly available in the Appendix.

2 The models

Assume $Y_1, ..., Y_n$ is a sample of results of an outcome of interest, and that for each subject i we also observe a vector of factors $Z_i = (X_i, M_i, W_i)$ where X_i, M_i and W_i are an exposure of interest, a mediator and a vector of confounders, respectively. We assume the conditional mean function for the outcome is $E(Y_i|Z_i) = g^{-1}(Z_i^T\beta)$

with g being the *link function* and where β is an unknown parameter vector. A consistent estimator, $\hat{\beta}$, for β is obtained as the solution to the estimating equations

$$U(\beta) = \sum_{i=1}^{n} D_i v_i^{-1}[y_i - E(Y_i|Z_i)] = 0 \tag{1}$$

where $D_i = \partial E(Y_i|Z_i)/\partial\beta$ and v_i is the working variance of y_i. By GEE theory, the variance of $\hat{\beta}$ can be consistently estimated by the robust sandwich estimator [26, 27].

In this paper, we consider a mediator, M, which may be binary, multilevel, continuous or a vector of any one of these. For convenience, we henceforth treat M as a scalar. Nevertheless, our framework also applies to the investigation of the joint mediation effect of multiple covariates. Consider the following conditional and marginal mean models for Y, with respect to M

$$E(Y|X, M, W) = g^{-1}(\beta_0 + \beta_1 X + \beta_2 M + \beta_3^T W) \tag{2}$$

$$E(Y|X, W) = g^{-1}(\beta_0^\star + \beta_1^\star X + \beta_3^{\star T} W). \tag{3}$$

Let $\mathscr{B} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\mathscr{B}^\star = (\beta_0^\star, \beta_1^\star, \beta_3^\star)$ be the vectors of conditional and marginal regression model parameters, and denote $\hat{\mathscr{B}}$ and $\hat{\mathscr{B}}^\star$ for their estimators obtained by solving eq. (1) under models (2) and (3), separately, respectively. When the two models (2) and (3) both hold simultaneously, we say we have *g-linkability*.

The definitions of the NDE and NIE, as given by [7], use the counterfactual framework. Let $Y(x, m)$ be the counterfactual outcome value when setting $X = x$ and $M = m$, and let $M(x)$ be the counterfactual mediator value when setting $X = x$. When comparing two exposure or treatment levels x and x', the TE, NDE and NIE can be then written [6, 7] as

$$\begin{aligned} TE(x, x') &= E(Y(x')) - E(Y(x)) = E(Y(x', M(x')) - E(Y(x, M(x))) \\ NDE(x, x') &= E(Y(x', M(x))) - E(Y(x, M(x))) \\ NIE(x, x') &= E(Y(x', M(x'))) - E(Y(x', M(x))). \end{aligned}$$

Under certain identifiability conditions, to estimate these effects, the *mediation formula*, given by [28, 29] and [6], can be used. Extensive work has been published on nonparametric and non-linear models [e.g., 6, 20, 21].

In GLMs, alternative definitions for the NDE, NIE and TE have been proposed. For example, for a binary outcome following a logistic model, [30] defined the TE on the odds ratio scale as

$$OR^{TE}(x, x'|W) = \frac{P(Y(x') = 1|W)/(1 - P(Y(x') = 1|W))}{P(Y(x) = 1|W)/(1 - P(Y(x) = 1|W))},$$

and have shown that it can be decomposed to a product of a NDE and a NIE. Therefore, on the log odds ratio scale, the TE decomposes to the sum of the NDE and the NIE. See [30] and [31] for further details. While these definitions depend on the value of the confounders, they show that, for a rare outcome, under the logistic regression outcome model and a linear mediator-exposure regression model, estimates of the NDE and NIE can be obtained using estimates of the regression model parameters. This result also extends to the log link function in GLMs, without a rare outcome assumption.

In this paper, we consider the causal effects on the link function scale, and assuming no exposure-mediator interaction. That is, as discussed by [32], under the assumptions given below, the TE, NDE and NIE are

$$\begin{aligned} TE(x, x'|W) &= g[E(Y(x')|W)] - g[E(Y(x)|W)] \\ NDE(x, x'|W) &= g[E(Y(x', M(x))|W)] - g[E(Y(x, M(x))|W)] \\ NIE(x, x'|W) &= g[E(Y(x', M(x'))|W)] - g[E(Y(x', M(x))|W)]. \end{aligned} \tag{4}$$

Throughout this paper, we assume that, after adjusting for measured confounders, there is no unmeasured confounding of the estimates of the exposure-outcome relationship, the mediator-outcome relationship or the exposure-mediator relationship. We also assume that confounders of the mediator-outcome relationship are unaffected by the exposure. Alternative identifiability assumptions have been given, e.g., in [6], which we will not consider further in this paper.

The mediation proportion, defined as NIE/TE, can be estimated using either the difference or the product method. The product method fits, in addition to model (2), a model for mediator-exposure relationship,

$$E(M|X, W) = \gamma_0 + \gamma_1 X + \gamma_2^T W. \tag{5}$$

Under the aforementioned no-confounding assumptions, considering a unit change (i.e., $x' - x = 1$) and when models (2) and (5) hold, the NIE is $\gamma_1\beta_2$, the NDE is β_1 and the mediation proportion is $\gamma_1\beta_2/(\gamma_1\beta_2 + \beta_1)$. The product method estimator for the mediation proportion is obtained by plugging in estimates of β_2 and γ_1 in this expression. Variance estimation and confidence intervals construction have been typically conducted by the bootstrap, due to the skewness of the finite sample distribution [2, 4].

In this paper, we focus on the difference method, for which we will develop asymptotic properties. Under g-linkability, the TE equals $\beta_1^\star$ and the NIE equals $\beta_1^\star - \beta_1$. Therefore, the mediation proportion equals to

$$p = \frac{\beta_1^\star - \beta_1}{\beta_1^\star} = 1 - \frac{\beta_1}{\beta_1^\star}.$$

If $p \in (0, 1]$, p can be interpreted as a proportion. The situation where $p = 0$ corresponds to $\beta_1 = \beta_1^\star$, hence in this case M does not mediate the effect of X at all. On the other hand, if $p = 1$ then the effect of X is fully mediated by M. Finally, if $p \notin [0, 1]$, the NDE and NIE are in opposite directions. One can get a consistent estimate for p by simply plugging in the appropriate estimator from each model. That is, $\hat{p} = 1 - \frac{\hat{\beta}_1}{\hat{\beta}_1^\star}$, where $\hat{\beta}_1$ and $\hat{\beta}_1^\star$ are the appropriate components of $\hat{\mathscr{B}}$ and $\hat{\mathscr{B}}^\star$. Under g-linkability, this estimator is consistent by standard GEE theory and the general mapping theorem.

The question of mediation can also be investigated when the available data is survival data. A counterfactual framework for mediation analysis of survival data has been previously provided [33–36]. [15] considered this question for the Cox model in the context of the PTE. First, as in [15], we define the following two models for the hazard function at time t, $h(t)$, conditionally and marginally, with respect to M

$$\begin{aligned} h(t|X, M, W) &= \lambda_0(t)\exp(\beta_0 + \beta_1 X + \beta_2 M + \beta_3^T W) \\ h(t|X, W) &= \lambda_0^\star(t)\exp(\beta_0^\star + \beta_1^\star X + \beta_3^{\star T} W), \end{aligned} \tag{6}$$

where $\lambda_0(t)$ and $\lambda_0^\star(t)$ are baseline hazard functions. [15] have shown that these two models cannot hold at same time. However, they claimed that if either $\beta_3^\star$ or $\Lambda_0^\star(t) = \int_0^t \lambda_0^\star(s)ds$ are small, then model (6) is a good approximation to the true conditional model. The assumption that $\Lambda_0^\star(t)$ is small is the rare outcome assumption. They confirmed this claim using a small scale simulation study. When (6) holds, approximately, the Cox model is approximately g-linkable. Thus, in addition to GLMs, we investigate in this paper estimation and inference for p in approximately g-linkable Cox models.

3 Further results on *g*-linkability

In this section, we consider the issue of when the full model (2) and the marginal model (3) both hold with the same function g exactly or approximately. Recall that g-linkability is sufficient for ensuring that $\hat{p}$, the point estimator of p, is consistent. This subject was also discussed in the context of random effects models [37], in which the authors showed, for each common statistical model, what random effect's distribution would provide a g-linkable conditional mean model condition on, and marginal over, the random effect. If

g-linkability does not hold, then $\hat{p}$ converges to $\overline{p} \neq p$. However, if g-linkability holds approximately, as in the case of logistic regression under rare outcome assumption (see below, and [38]), then one may expect $\overline{p}$ to be close to p, as discussed in [15] for the Cox model.

We consider the three common link functions: identity, log and logit. For each of these functions we give a general condition for the distribution of M given X and W that ensures g-linkability, where in the logit link function a rare outcome assumption is also needed. Numerous detailed and practical examples that fulfill these conditions can be constructed. In practice, the difference method does not require fitting of the mediator-exposure relationship model, as noted by [39]. For the validity of the product method, however, this model has to be correctly specified.

3.1 Identity link function

Under the identity link function, models (2) and (3) simplify to

$$E(Y|X, M, W) = \beta_0 + \beta_1 X + \beta_2 M + \beta_3^T W$$
$$E(Y|X, W) = \beta_0^\star + \beta_1^\star X + {\beta_3^\star}^T W.$$

We now show that g-linkability holds whenever $E(M|X, W)$ is a linear function of X and W. To see that, let $E(M|X, W) = a + b_1 X + b_3^T W$, for some a, b_1 and b_3. Then,

$$\begin{aligned} E(Y|X, W) = E(E(Y|X, M, W)|X, W) &= \beta_0 + \beta_1 X + \beta_2 E(M|X, W) + \beta_3^T W \\ &= \beta_0^\star + \beta_1^\star X + {\beta_3^\star}^T W \end{aligned}$$

where $\beta_0^\star = \beta_0 + \beta_2 a$, $\beta_1^\star = \beta_1 + \beta_2 b_1$ and $\beta_3^\star = \beta_3 + \beta_2 b_3$.

3.2 Log link function

Under the log link function, $g(u) = \log(u)$, the mean models become

$$E(Y|X, M, W) = \exp(\beta_0 + \beta_1 X + \beta_2 M + \beta_3^T W)$$
$$E(Y|X, W) = \exp(\beta_0^\star + \beta_1^\star X + {\beta_3^\star}^T W)$$

and we have

$$E(Y|X, W) = E(E(Y|X, M, W)|X, W) = \exp(\beta_0 + \beta_1 X + \beta_3^T W) \times E[\exp(\beta_2 M)|X, W].$$

Therefore, in the log link case, g–linkability holds if the log of the moment generating function of $M|X, W$ can be written as a linear function of X and W. That is, $\log E[\exp(\beta_2 M)|X, W] = a' + b_1' X + {b_3'}^T W$.

3.3 Logit link function

The issue of whether the logistic regression model holds for both the conditional and marginal models has been discussed in the literature [37–39]. The logit link function, defined as $\text{logit(p)} = \log(\text{p}/(1-\text{p}))$, is typically used when Y is binary. It is well known and readily seen that when the outcome is rare, the logit function is similar to the log function. Thus, under rare outcome scenario, one may expect that g-linkability holds approximately for the logit link function, as is typical in many epidemiologic and public health studies [30]. We empirically investigate the limits of the rare outcome assumption in Section 5.

4 Inference for the mediation proportion

For simplicity of presentation, we assume throughout this section that g-linkability holds. Then, $\hat{p} = 1 - \frac{\hat{\beta}_1}{\hat{\beta}_1^\star}$, where $\hat{\beta}_1$ and $\hat{\beta}_1^\star$ are the appropriate components of $\hat{\mathscr{B}}$ and $\hat{\mathscr{B}}^\star$ defined in Section 2. By the aforementioned GEE theory together with the general mapping theorem, this estimator is consistent.

Asymptotic confidence intervals for p have been constructed previously using either Fieller's theorem or the delta method [14, 15, 40]. Here, we consider the latter. As written in [15], by the delta method $\hat{p}$ has an asymptotic normal distribution with variance equals to

$$\sigma^2_{\hat{p}} = \frac{\sigma^2_{\hat{\beta}_1}}{(\beta_1^\star)^2} + \frac{\beta_1^2 \sigma^2_{\hat{\beta}_1^\star}}{(\beta_1^\star)^4} - 2\frac{\beta_1 \sigma_{\hat{\beta}_1, \hat{\beta}_1^\star}}{(\beta_1^\star)^3}, \tag{7}$$

where

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1), \quad \sigma^2_{\hat{\beta}_1^\star} = Var(\hat{\beta}_1^\star) \quad \text{and} \quad \sigma_{\hat{\beta}_1, \hat{\beta}_1^\star} = Cov(\hat{\beta}_1, \hat{\beta}_1^\star).$$

While $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_1^\star}$ can be consistently estimated using the robust sandwich estimator [26] for each of the models (2) and (3) separately, it is not obvious how to estimate $\sigma_{\hat{\beta}_1, \hat{\beta}_1^\star}$, the covariance of estimators obtained from two separate models. In Section 4.1, we propose a data duplication algorithm to estimate this quantity.

Assume now we have estimates $\hat{\sigma}^2_{\hat{\beta}_1}$, $\hat{\sigma}^2_{\hat{\beta}_1^\star}$ and $\hat{\sigma}_{\hat{\beta}_1, \hat{\beta}_1^\star}$ for $\sigma^2_{\hat{\beta}_1}$, $\sigma^2_{\hat{\beta}_1^\star}$ and $\sigma_{\hat{\beta}_1, \hat{\beta}_1^\star}$, respectively. These estimates are plugged in (7) in order to get an estimate $\hat{\sigma}^2_{\hat{p}}$ and a $(1-\alpha)$ level confidence interval for p may be obtained as

$$\hat{p} \pm z_{1-\alpha/2} \hat{\sigma}_{\hat{p}} \tag{8}$$

with $z_{1-\alpha/2}$ being the appropriate quantile of the normal distribution.

Past authors concentrated on methods for testing that the mediation proportion is at least some fraction f, with f typically being 0.5 or more [14, 15, 40]. In the context of PTE, where the validation of intermediate biomarkers for outcome is of interest, this may be reasonable. However, when considering a mediator, the more relevant question is whether M is indeed a mediator. Then, the hypothesis is $H_0 : p = 0$ vs. $H_1 : p \neq 0$. Let $Z_p = \sigma_p^{-1} \hat{p}$ be the scaled estimate. By the delta method, the distribution of Z_p under the null converges to a standard normal distribution and the null is rejected at significance level α if $|Z_p| > z_{1-\alpha/2}$. An alternative test statistic is based upon a test for the difference between the effect estimates in the marginal and conditional models. That is, on $\hat{d} = \hat{\beta}_1^\star - \hat{\beta}_1$. Under the assumptions in this paper, $\hat{d}$ is a consistent estimate for the NIE. A test statistic based on $\hat{d}$ is based on $Z_d = \sigma_{\hat{d}}^{-1} \hat{d}$, where

$$\sigma^2_{\hat{d}} = Var(\hat{d}) = \sigma^2_{\hat{\beta}_1} + \sigma^2_{\hat{\beta}_1^\star} - 2\sigma_{\hat{\beta}_1, \hat{\beta}_1^\star}.$$

Then, the null hypothesis is rejected if $|Z_d| > z_{1-\alpha/2}$.

4.1 The data duplication algorithm

A main challenge when conducting inference for the mediation proportion p is to estimate the covariance of estimators obtained from the two models (2) and (3). It turns out that the covariance between $\hat{\mathscr{B}}$ and $\hat{\mathscr{B}}^\star$ can be estimated by fitting both models by stacking the estimating equations for the two models using a data duplication algorithm. A similar method was presented in [15] for the Cox model in survival data. Here, we extend it to GEE for GLMs. First, the data are augmented with additional pseudo-variables and pseudo-observations. Each variable, including the intercept, the exposure, the confounders, but not the mediator,

Table 1: The augmented data used by the data duplication algorithm. For each original observation i, two rows $j = 1, 2$ are created. The duplicated data is used for the pseudo model presented in eq. (9).

i	j	Intercept	Intercept*	X	X*	M	W	W*	Y
1	1	1	0	x_1	0	m_1	w_1	0	y_1
1	2	0	1	0	x_1	0	0	w_1	y_1
2	1	1	0	x_2	0	m_2	w_2	0	y_2
2	2	0	1	0	x_2	0	0	w_2	y_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

appears twice, and each of the original observations is included as two pseudo-observations in the new data set. See Table 1 for an illustration of the duplicated data structure.

The following pseudo model is fitted to the duplicated data using GEE [27],

$$\begin{aligned} E(Y_{ij}|X_i, X_i^\star, M_i, W_i, W_i^\star) = g^{-1}(&\beta_0 I\{j=1\} + \beta_1 X_i + \beta_2 M_i + \beta_3^T W_i \\ &+ \beta_0^\star I\{j=2\} + \beta_1^\star X_i^\star + \beta_3^T W_i^\star), \end{aligned} \tag{9}$$

where $j = 1, 2$ are the rows created from duplicating each observation and are treated as repeated measures. Model (9) implies that we can write $E(Y_{i1}|X_i, X_i^\star, M_i, W_i, W_i^\star) = E(Y_{i1}|X_i, M_i, W_i)$ and $E(Y_{i2}|X_i, X_i^\star, M_i, W_i, W_i^\star) = E(Y_{i2}|X_i^\star, W_i^\star)$. Let R be a 2×2 working correlation matrix and denote $B_i = diag(v_{i1}, v_{i2})$, where $v_{ij} = Var(Y_{ij})$. Let also $V_i = B_i^{1/2} R B_i^{1/2}$ be a 2×2 working variance for the vector (Y_{i1}, Y_{i2}). Here, the GEE are defined as

$$U_{GEE}(\mathscr{B}) = \sum\nolimits_{i=1}^{n} (\mathscr{D}_i, \mathscr{D}_i^\star) V_i^{-1} \begin{pmatrix} y_{i1} - E(Y_{i1}|X_i, M_i, W_i) \\ y_{i2} - E(Y_{i2}|X_i^\star, W_i^\star) \end{pmatrix}, \tag{10}$$

where $\mathscr{D}_i = \partial E(Y_{i1}|X_i, M_i, W_i)/\partial \mathscr{B}$ and $\mathscr{D}_i^\star = \partial E(Y_{i2}|X_i^\star, W_i^\star)/\partial \mathscr{B}^\star$ are two column vectors. If R is taken to be the identity matrix, then $V_i = B_i$ and eq. (10) simplifies to the following estimating equations

$$U_{IEE}(\mathscr{B}) = \begin{pmatrix} U_{IEE}^{(1)}(\beta) \\ U_{IEE}^{(2)}(\beta^\star) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \mathscr{D}_i v_{i1}^{-1} [y_i - E(Y_{i1}|X_i, M_i, W_i)] \\ \sum_{i=1}^{n} \mathscr{D}_i^\star v_{i2}^{-1} [y_i - E(Y_{i2}|X_i^\star, W_i^\star)] \end{pmatrix} = 0. \tag{11}$$

Then, the estimating equations given by eq. (11) are identical to the estimating equations for fitting models (2) and (3) separately, because D_i, v_i and Z_i in eq. (1) are equal to $\mathscr{D}_i$, v_{i1} and (X_i, M_i, W_i), respectively, under model (2), and they are equal to $\mathscr{D}_i^\star$, v_{i2} and $(X_i^\star, W_i^\star)$, respectively, under model (3). The major advantage of the data duplication algorithm is that it provides an estimator for $\sigma_{\beta_1, \beta_1^\star}$ in a straightforward manner. Taking a working correlation matrix other than the identity may result in more efficient estimators of $\hat{\mathscr{B}}$, but would not have the desirable property that the duplicated data estimating equations are identical to the two separate estimating equations from the two separate models.

5 Simulation study

The simulation studies and data analysis were conducted in **R**. The code is available upon request from the first author. In addition, we have developed a **SAS** macro and an **R** package that are publicly available (Appendix A.1). In the simulation studies, we considered several issues regarding the performance of the methodology we presented throughout the paper. We first present results concerning g-linkability for the logit link function and the Cox model. Then, we turn to the performance of the mediation proportion estimator, studying its bias, the coverage rate of the accompanied confidence interval and the type I error and the power of the statistical tests described in Section 4.

Throughout these simulation studies, we assume that there are no confounders in the model. X and M were generated using a bivariate normal with mean $(0,0)^T$ and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Then, we have that $\beta_2 = \frac{p}{\rho}\beta_1^\star$ for the identity, log and logit link functions (the latter under the rare outcome assumption); see Web Appendix A. In these scenarios, g–linkability holds for all three link functions. The estimation and inference procedures apply to any bivariate distribution of X and M that satisfies the simple moment conditions given in Section 3, and here we used the bivariate normal distribution for generating the data merely for convenience. The estimation and inference procedures do not use the bivariate normal distribution of (X, M).

5.1 *g*-linkability for the logit link function and of the Cox model

In order to assess the magnitude of the bias when assuming g-linkability of the logit link function and the Cox model, we conducted a simulation study under various conditions and inspected the resulting bias in $\hat{p}$, as estimated using the data duplication algorithm described in Section 4.1 while taking the working correlation matrix to be the identity. First we describe the logit link function model. We simulate Y under the logistic regression model

$$\text{logit}(\text{P}(\text{Y} = 1|\text{X}, \text{M})) = \beta_0 + \beta_1\text{X} + \beta_2\text{M}.$$

We chose the model parameter values in the following way. First, we chose $\rho = corr(X, M)$, p and $\beta_1^\star$. Then, by definition we had $\beta_1 = (1-p)\beta_1^\star$, and we took β_2 as if g-linkability exactly holds. That is, $\beta_2 = \frac{p}{\rho}\beta_1^\star$. Then, we fixed the unconditional case probability $P(Y = 1)$ and found the appropriate β_0 value by solving for β_0 in the equation

$$P(Y = 1) = E(\text{expit}(\beta_0 + \beta_1\text{X} + \beta_2\text{M})),$$

where $\text{expit(u)} = \text{exp(u)}/(1 + \text{exp(u)})$. Finally, the sample size was given as $n = E(N_{cases})/P(Y = 1)$ where $E(N_{cases})$ is number of expected cases. We considered the following values for the parameters. $p = 0.1, 0.2, \ldots, 0.8$; $\rho = p, p + 0.1, \ldots, 0.8$, with $\rho \geq p$ to satisfy that $\beta_2 \leq \beta_1^\star$ or in words, to ensure that the total effect of X is larger than effect of M; $\beta_1^\star = \log(1.25), \log(1.5), \log(2)$; $P(Y = 1) = 0.005, 0.01, 0.1, 0.25$; $E(N_{cases}) = 100, 500, 1000$. The number of simulation iterations per scenario was 1000.

For the Cox model, we simulated the data similarly to the logit link function simulations. First, we simulated X and M as before. Then, given fixed ρ, p and $\beta_1^\star$, $\beta_2 = \frac{p}{\rho}\beta_1^\star$. We took a Weibull distribution for the baseline hazard and used Exponential distribution for the censoring (mean=50), with additional cutoff at age 90. Given the desired proportion number of cases in the population, we used simulations to find the appropriate values for the Weibull distribution shape parameter, while fixing the scale parameter at 200. As in the logit link case, we chose the sample size as the number of expected cases ($E(N_{cases})$) divided by the expected proportion of cases ($P(\delta = 1)$), where δ is the event indicator.

In order to assess g-linkability, and the finite sample performance of $\hat{p}$, we calculated the relative bias, defined as $100 \times |\frac{mean(\hat{p})-p}{p}|$. Ideally, this quantity should be close to zero. We note that bias may arise either because g-linkability fails to hold, or because of a sample size not large enough. Figure 1 presents bias for $\beta_1^\star = \log(1.5)$ as a function of the parameters. First, it is of note that whenever the overall prevalence or cumulative incidence of Y was small, as in the rare disease scenario, and the number of cases was sufficiently large, bias was minimal. Even when the disease was not as rare, e.g., $P(Y = 1) = 0.25$, when there were enough cases, and when p was large enough (e.g., $p > 0.2$ in this case), the bias was minimal. Considering the g-linkability of the Cox model, presented for $\beta_1^\star = \log(1.5)$ in Figure 2, the results were similar to the results obtained for the logit link function. That is, when the outcome was rare ($P(\delta = 1)$ was small) then the corresponding marginal Cox model for the hazard function approximately holds and the bias in mediation proportion estimation was minimal. Figures similar to Figures 1 and 2 are presented for $\beta_1^\star = \log(1.25), \log(2)$ in Web Appendix B. The overall trends were similar.

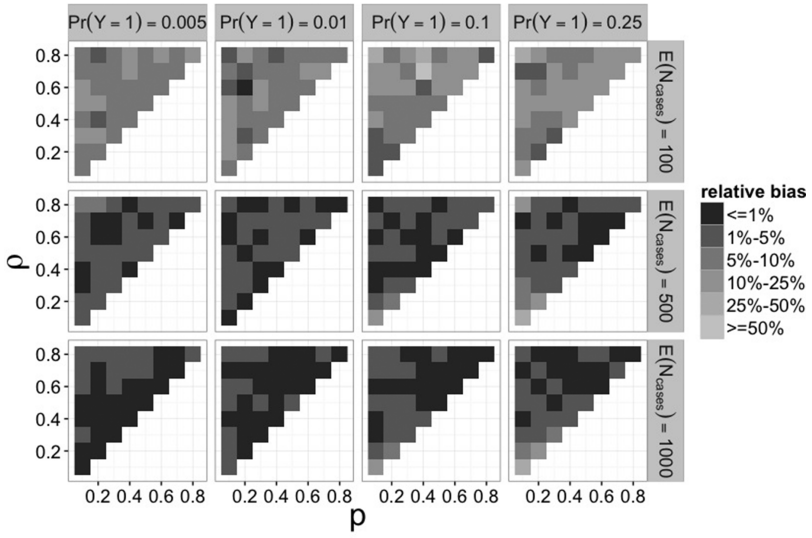

Figure 1: Relative bias of the mediation proportion estimator under the logistic model as a function of the mediation proportion (p), the correlation between the exposure and the mediator (ρ), the number of expected cases ($E(N_{cases})$) and the outcome rate ($P(Y = 1)$). The value of $\beta_1^\star$ was taken to be log(1.5).

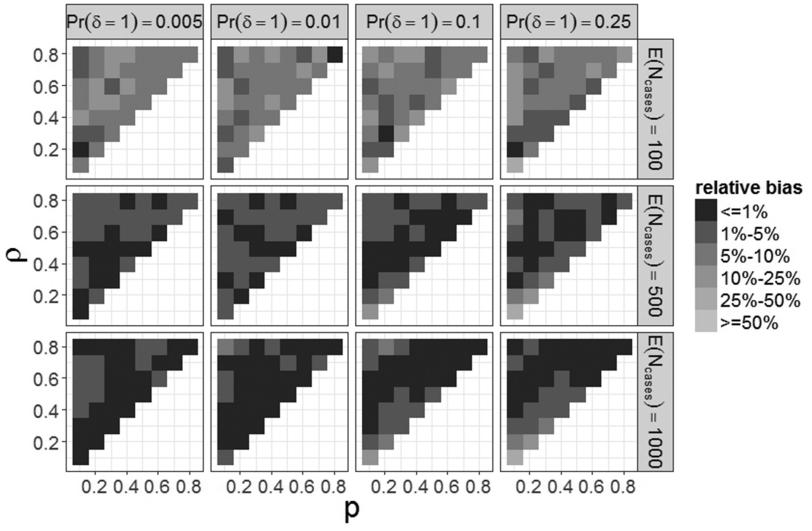

Figure 2: Relative bias of the mediation proportion estimator under the Cox model as a function of the mediation proportion (p), the correlation between the exposure and the mediator (ρ), the number of expected cases ($E(N_{cases})$) and the event rate ($P(\delta = 1)$). The value of $\beta_1^\star$ was taken to be log(1.5).

5.2 Estimation and inference performance

For the Cox model and the logit link function, data were simulated as described above. For the identity link function, data were simulated from the model $Y = E(Y|X, M) + \epsilon = \beta_0 + \beta_1 X + \beta_2 M + \epsilon$. As before, (X, M) were simulated from a bivariate normal distribution with zero mean, unit variance, and correlation ρ. The error, ϵ, was a vector of iid standard normal random variables. We also considered other distributions for ϵ; we will expand on this matter later on. As in the logit link case, we fixed $\beta_1^\star$, ρ, and p and took $\beta_1 = (1 - p)\beta_1^\star$ and $\beta_2 = \frac{p}{\rho}\beta_1^\star$. The intercept, β_0, was chosen arbitrarily to be equal to 2. We considered various values for $\beta_1^\star$, p and ρ, where as before we were only interested in scenarios where $\rho \geq p$, since then $\beta_2 \geq \beta_1^\star$. We present

Table 2: Percent relative bias and efficiency of estimators of the mediation proportion under the logistic link function. *RBiasDiff* is the difference between the relative bias of difference method compared to that of the product method times 100. Negative values indicate that the difference method is preferable. *Vratio* is the ratio of the empirical variances of the estimates of the difference and product methods. *Vratio* > 1 indicates that the product method is preferable.

$P(Y=1)$	p	ρ	$\exp(\beta_1^\star)=1.25$		$\exp(\beta_1^\star)=1.5$		$\exp(\beta_1^\star)=2$	
			%RBiasDiff†	*Vratio*	*%RBiasDiff*†	*Vratio*	*%RBiasDiff*†	*Vratio*
0.005	0.1	0.1	−0.23	1.00	−0.82	1.00	2.84	1.14
		0.5	0.01	1.00	0.04	1.00	−0.14	1.00
		0.7	−0.01	1.00	0.02	1.00	−0.04	1.00
	0.3	0.3	−0.07	1.00	−0.20	1.00	0.12	0.99
		0.5	−0.02	1.00	−0.07	1.00	−0.26	0.99
		0.7	−0.00	1.00	−0.02	1.00	−0.10	1.00
	0.5	0.5	−0.02	1.00	−0.08	1.00	−0.33	1.00
		0.7	−0.01	1.00	−0.02	1.00	−0.11	1.00
0.01	0.1	0.1	−0.55	0.99	−1.00	1.02	6.00	1.13
		0.5	0.02	1.00	−0.08	1.00	0.22	0.99
		0.7	0.02	1.00	0.04	1.00	−0.12	1.00
	0.3	0.3	−0.11	1.00	−0.42	0.99	0.88	1.01
		0.5	−0.03	1.00	−0.12	1.00	−0.04	0.99
		0.7	−0.01	1.00	−0.04	1.00	0.15	1.00
	0.5	0.5	−0.04	1.00	−0.14	1.00	−0.35	1.00
		0.7	−0.01	1.00	−0.04	1.00	−0.20	1.00
0.1	0.1	0.1	−1.44	0.97	11.52	0.97	46.42	1.33
		0.5	0.27	1.00	−0.63	0.99	1.45	0.96
		0.7	−0.18	1.00	−0.32	1.00	−0.24	0.98
	0.3	0.3	−0.93	0.99	−0.86	0.97	11.26	0.97
		0.5	−0.30	1.00	−1.05	0.98	2.99	0.94
		0.7	−0.10	1.00	−0.36	0.99	−1.20	0.97
	0.5	0.5	−0.30	1.00	−0.35	1.00	2.71	1.01
		0.7	−0.09	1.00	−0.41	1.00	0.62	0.99

†$\%Rbiasdiff = 100 \times \left[\left|\frac{Mean(\hat{p}_{Diff})-p}{p}\right| - \left|\frac{Mean(\hat{p}_{Prod})-p}{p}\right|\right]$

results for $\beta_1^\star = 0.1, 0.3, 0.5$, which imply multiple correlations between (X, M) and Y of about 0.1, 0.3 and 0.5, respectively.

Under the simple linear model (identity link), estimates obtained from the difference and product methods are algebraically identical [41]. This result does not extend to the logistic link function or to the Cox model [3, 42]. We therefore compared between the difference and the product methods. Let $\hat{p}_{diff}$ and $\hat{p}_{prod}$ be the difference and product method estimators, respectively. Table 2 presents $\%RBiasDiff = 100 \times [|\frac{mean(\hat{p}_{diff})-p}{p}| - |\frac{mean(\hat{p}_{prod})-p}{p}|]$, and the empirical variance ratio $\frac{Var(\hat{p}_{diff})}{Var(\hat{p}_{ratio})}$. The two methods are generally comparable when g-linkability holds.

From estimation we move to hypothesis testing. The two test statistics compared were described in Section 4, where the variance estimators used in the test statistics were obtained by the data duplication algorithm described in Section 4.1. Results are presented in Table 3. In terms of type I error, both tests were adequate, with a conservative type I error when the correlation between the exposure and the mediator was low. When the total effect was low, the test based on $\hat{d}$ had greater power, usually by 5% – 10%, compared to the test based on $\hat{p}$. The power of both tests was highly affected by the effect size ($\beta_1^\star$) and the correlation between the exposure and the mediator (ρ). High correlation between X and M decreased the power. The power of both tests was lower when the total effect $\beta_1^\star$ is low. It should be noted that mediation analysis is performed only after a risk factor was found to be significant, which, in general, is less likely to happen if both $\beta_1^\star$ and the sample size are low. We further address this point in Section 7. We next consider the finite sample properties of the confidence interval presented in Section 4. Table 4 presents empirical coverage rates and confidence interval widths. Coverage rates were generally adequate. When ρ and p were of similar size,

Table 3: Type *I* error and power for tests for mediation under the identity ($n = 1000$) and logit link functions and the Cox model ($E(N_{cases}) = 500$).

						Identity link function ($Y \sim Normal$)	
		$\beta_1^\star = 0.1$		$\beta_1^\star = 0.3$		$\beta_1^\star = 0.5$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.00	0.02	0.01	0.01	0.01	0.02
	$\rho = 0.5$	0.01	0.04	0.05	0.05	0.05	0.05
	$\rho = 0.7$	0.01	0.05	0.04	0.05	0.06	0.06
$p = 0.1$	$\rho = 0.1$	0.28	0.60	0.91	0.90	0.90	0.88
	$\rho = 0.5$	0.02	0.07	0.37	0.38	0.79	0.79
	$\rho = 0.7$	0.02	0.05	0.14	0.15	0.35	0.36
$p = 0.2$	$\rho = 0.3$	0.28	0.52	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.07	0.18	0.90	0.91	1.00	1.00
	$\rho = 0.7$	0.03	0.10	0.47	0.49	0.89	0.89
$p = 0.3$	$\rho = 0.3$	0.54	0.83	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.17	0.36	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.06	0.18	0.83	0.84	0.99	1.00
						Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.01$	
		$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$		$\beta_1^\star = \log(2)$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.03	0.06	0.04	0.05	0.04	0.04
	$\rho = 0.5$	0.03	0.05	0.05	0.06	0.05	0.05
	$\rho = 0.7$	0.04	0.05	0.04	0.04	0.05	0.05
$p = 0.1$	$\rho = 0.1$	1.00	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.09	0.13	0.34	0.36	0.74	0.74
	$\rho = 0.7$	0.06	0.09	0.15	0.16	0.35	0.36
$p = 0.2$	$\rho = 0.3$	0.85	0.89	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.31	0.40	0.86	0.87	1.00	1.00
	$\rho = 0.7$	0.13	0.17	0.42	0.44	0.87	0.87
$p = 0.3$	$\rho = 0.3$	0.99	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.66	0.73	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.25	0.32	0.74	0.76	0.99	0.99
						Cox model with $P(\delta = 1) = 0.01$	
		$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$		$\beta_1^\star = \log(2)$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.03	0.05	0.04	0.04	0.04	0.04
	$\rho = 0.5$	0.03	0.05	0.04	0.05	0.06	0.06
	$\rho = 0.7$	0.03	0.05	0.06	0.06	0.06	0.06
$p = 0.1$	$\rho = 0.1$	0.99	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.09	0.14	0.33	0.35	0.75	0.76
	$\rho = 0.7$	0.05	0.09	0.15	0.16	0.32	0.33
$p = 0.2$	$\rho = 0.3$	0.84	0.88	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.33	0.43	0.85	0.86	1.00	1.00
	$\rho = 0.7$	0.15	0.19	0.46	0.47	0.87	0.88
$p = 0.3$	$\rho = 0.3$	0.99	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.67	0.75	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.25	0.33	0.78	0.80	1.00	1.00

and the total effect was large, the new method did not produce confidence intervals with nominal coverage. For the logistic link function and the Cox model, the worst results were obtained for the combination of $p = \rho = 0.1$ and $\log(\beta_1^\star) = 2$. As can be seen from Figures 1 and 2 (and from the figures in Web Appendix B) in these cases, *g*-linkability does not hold because the outcome is not that rare and the relative risk is strong.

For the difference method, we also compared the confidence intervals using the asymptotic variance to confidence intervals constructed using the bootstrap, both by estimating the bootstrapped variance and

Table 4: Empirical coverage rates (CR) and lengths (LEN) of confidence intervals for the mediation proportion under the identity and logit link functions and the Cox model

						Identity link function ($Y \sim Normal$)		
				$n = 1000$				$n = 5000$
			$\beta_1^\star = 0.1$	$\beta_1^\star = 0.3$	$\beta_1^\star = 0.5$	$\beta_1^\star = 0.1$	$\beta_1^\star = 0.3$	$\beta_1^\star = 0.5$
$p = 0.1$	$\rho = 0.1$	CR	0.90	0.96	0.95	0.94	0.95	0.94
		LEN	0.24	0.13	0.12	0.10	0.06	0.05
	$\rho = 0.5$	CR	0.98	0.95	0.96	0.96	0.95	0.96
		LEN	0.85	0.25	0.15	0.33	0.11	0.07
	$\rho = 0.7$	CR	0.98	0.96	0.95	0.96	0.95	0.96
		LEN	1.42	0.41	0.24	0.56	0.18	0.11
$p = 0.3$	$\rho = 0.3$	CP	0.93	0.95	0.94	0.96	0.97	0.95
		LEN	0.65	0.20	0.14	0.25	0.09	0.06
	$\rho = 0.5$	CR	0.96	0.96	0.94	0.96	0.95	0.96
		LEN	0.93	0.28	0.17	0.37	0.12	0.07
	$\rho = 0.7$	CR	0.97	0.96	0.96	0.95	0.95	0.97
		LEN	1.51	0.43	0.26	0.59	0.19	0.11
$p = 0.5$	$\rho = 0.5$	CR	0.94	0.94	0.96	0.96	0.96	0.95
		LEN	1.11	0.32	0.20	0.43	0.14	0.09
	$\rho = 0.7$	CR	0.96	0.95	0.95	0.96	0.96	0.95
		LEN	1.65	0.47	0.27	0.63	0.20	0.12
						Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.01$		
					$E(N_{cases}) = 500$			$E(N_{cases}) = 1000$
					$\beta_1^\star$			$\beta_1^\star$
			log(1.25)	log(1.5)	log(2)	log(1.25)	log(1.5)	log(2)
$p = 0.1$	$\rho = 0.1$	CR	0.95	0.95	0.87	0.96	0.93	0.80
		LEN	0.13	0.07	0.04	0.08	0.05	0.03
	$\rho = 0.5$	CR	0.97	0.96	0.94	0.96	0.95	0.96
		LEN	0.49	0.26	0.15	0.33	0.19	0.11
	$\rho = 0.7$	CR	0.96	0.95	0.96	0.95	0.94	0.95
		LEN	0.84	0.43	0.25	0.57	0.31	0.18
$p = 0.3$	$\rho = 0.3$	CR	0.96	0.95	0.92	0.95	0.95	0.94
		LEN	0.39	0.20	0.11	0.26	0.13	0.08
	$\rho = 0.5$	CR	0.96	0.95	0.93	0.96	0.96	0.95
		LEN	0.55	0.29	0.17	0.37	0.20	0.12
	$\rho = 0.7$	CR	0.96	0.94	0.96	0.95	0.95	0.94
		LEN	0.88	0.47	0.26	0.59	0.32	0.19
$p = 0.5$	$\rho = 0.5$	CR	0.95	0.95	0.94	0.96	0.96	0.94
		LEN	0.65	0.34	0.20	0.45	0.24	0.14
	$\rho = 0.7$	CR	0.97	0.94	0.95	0.96	0.95	0.93
		LEN	0.95	0.49	0.29	0.66	0.34	0.20
						Cox model with $P(\delta = 1) = 0.01$		
					$E(N_{cases}) = 500$			$E(N_{cases}) = 1000$
					$\beta_1^\star$			$\beta_1^\star$
			log(1.25)	log(1.5)	log(2)	log(1.25)	log(1.5)	log(2)
$p = 0.1$	$\rho = 0.1$	CR	0.93	0.95	0.92	0.93	0.81	0.73
		LEN	0.12	0.08	0.07	0.05	0.06	0.04
	$\rho = 0.5$	CR	0.96	0.95	0.96	0.94	0.94	0.95
		LEN	0.48	0.34	0.26	0.18	0.15	0.10
	$\rho = 0.7$	CR	0.98	0.95	0.95	0.96	0.97	0.94
		LEN	0.83	0.56	0.43	0.30	0.25	0.18
$p = 0.3$	$\rho = 0.3$	CR	0.95	0.96	0.95	0.95	0.93	0.91
		LEN	0.38	0.26	0.19	0.13	0.11	0.08
	$\rho = 0.5$	CR	0.97	0.96	0.95	0.95	0.95	0.94
		LEN	0.57	0.38	0.29	0.20	0.17	0.12
	$\rho = 0.7$	CR	0.96	0.94	0.95	0.96	0.95	0.95
		LEN	0.87	0.60	0.45	0.32	0.26	0.18
$p = 0.5$	$\rho = 0.5$	CR	0.96	0.95	0.94	0.96	0.94	0.94
		LEN	0.66	0.44	0.34	0.24	0.20	0.14
	$\rho = 0.7$	CR	0.96	0.96	0.95	0.95	0.95	0.95
		LEN	0.94	0.63	0.49	0.34	0.28	0.20

Table 5: Ratio between mean estimated $\widehat{Cov}(\hat{\beta}_1^\star, \hat{\beta}_1)$ and empirical $Cov(\hat{\beta}_1^\star, \hat{\beta}_1)$ under the identity ($n = 1000$) and logit link functions and the Cox model $E(N_{cases}) = 100$

				Identity link function ($Y \sim Normal$)
		$\beta_1^\star = 0.1$	$\beta_1^\star = 0.3$	$\beta_1^\star = 0.5$
$p = 0.1$	$\rho = 0.1$	0.989	0.994	0.985
	$\rho = 0.5$	1.036	1.014	0.963
	$\rho = 0.7$	0.960	0.962	1.098
$p = 0.3$	$\rho = 0.3$	1.031	0.959	0.967
	$\rho = 0.5$	1.103	0.889	0.989
	$\rho = 0.7$	0.988	0.975	0.955
$p = 0.5$	$\rho = 0.5$	1.064	0.959	1.008
	$\rho = 0.7$	1.010	0.980	1.100
				Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.1$
		$\beta_1^\star = \log(1.25)$	$\beta_1^\star = \log(1.5)$	$\beta_1^\star = \log(2)$
$p = 0.1$	$\rho = 0.1$	0.996	0.968	1.031
	$\rho = 0.5$	1.070	0.977	1.023
	$\rho = 0.7$	1.012	0.976	0.940
$p = 0.3$	$\rho = 0.3$	1.002	0.954	1.057
	$\rho = 0.5$	0.954	1.016	1.080
	$\rho = 0.7$	0.958	0.902	0.995
$p = 0.5$	$\rho = 0.5$	0.993	0.983	1.008
	$\rho = 0.5$	0.996	1.011	0.959
				Cox model with $P(\delta = 1) = 0.1$
		$\beta_1^\star = \log(1.25)$	$\beta_1^\star = \log(1.5)$	$\beta_1^\star = \log(2)$
$p = 0.1$	$\rho = 0.1$	1.089	1.035	0.951
	$\rho = 0.5$	1.011	1.013	0.945
	$\rho = 0.7$	1.099	1.044	1.068
$p = 0.3$	$\rho = 0.3$	1.053	0.892	0.906
	$\rho = 0.5$	0.957	0.990	0.887
	$\rho = 0.7$	0.999	0.982	0.966
$p = 0.5$	$\rho = 0.5$	0.945	0.938	1.036
	$\rho = 0.5$	0.961	1.057	0.963

assuming normality, and by using the quantiles of the bootstrap samples. The results, presented in Web Appendix C, show that the asymptotic approach is comparable to both bootstrap procedures in terms of nominal coverage. Both versions of the bootstrap confidence intervals were wider than the asymptotic confidence intervals. Furthermore, the bootstrap is time consuming, especially for large data sets, where our method, implemented by publicly-available software, is as fast as a single GEE model fit. For $n = 5000$, the computation time of 500 bootstrap replications, for a single data set, was about 115 seconds; For $n = 10000$, it was 250 seconds, and for $n = 50000$ it was about 1300 seconds (more than 20 minutes). For comparison, calculation of confidence intervals in our method was less than one second for $n = 5000, 10000$ and about 4 seconds for $n = 50000$.

Throughout this section, we presented in parallel results for the identity and logit link function and the Cox model. There was a very strong agreement between the results for the logit link function for binary data and the Cox model, as one may have expect given the close relationship between the logistic regression model and the Cox model in epidemiology and public health evaluations.

In addition to the scenarios we described above, we conducted simulations for the identity link function with error distributions other than the normal one. We considered symmetric distribution with tails heavier than the normal distribution as well as skewed distributions. As predicted by GEE theory, the performance of the mediation proportion estimator, the statistical tests and the confidence interval was only slightly changed. Details are given in Web Appendix D.

6 Illustrative example

We illustrate the use of our methodology in an analysis of the etiology of pre-menopausal breast cancer data from the Nurses Health's Studies (NHS and NHSII) [43, 44]. It was previously found that high mammographic density (MD) is a risk factor for breast cancer [45]. The goal here is to investigate whether, and to what extent, the effects of more distal risk factors for pre-menopausal breast cancer are mediated by high MD. A detailed description of this study is given in [46]. In this nested case-control study, controls were matched to cases by current age, menopausal status, current hormone use, month, time of day, fasting status and time of the day at blood collection and luteal day (for NHSII samples only). There were 559 pre-menopausal cases and 1727 controls. Since the disease is rare, as shown in the previous section, g-linkability should hold. Mediation by percent MD, a single mediator, was considered for a number of well-established breast cancer risk factors. Following [46], the mediation analysis was conducted for each risk factor separately.

We only considered risk factors with significant total effects: personal history of benign breast disease (HBBD), family history of breast cancer (FH), adolescent somatotype (ASM), body mass index at age 18 (BMI18), age at first birth (AFB), age at menarche (AM) and height (HT). In each of the analyses, we included potential confounders for the risk factor-MD, MD-outcome and risk factor-outcome relationships, based upon subject matter considerations following [46]. For example, we did not adjust for current (adult) BMI in the analysis of BMI18, as the latter affects the former. In addition, some of the risk factors studied may have been confounders in analysis of mediation via MD of another risk factor. The set of confounders used in at least one analysis included current age, fasting status, blood collection time of the day, mammography batch (NHS batch 1, NHS batch 2 or NHSII), current BMI, BMI18, ASM, HBBD, parity, AFB, and AM. As in most observational studies, residual confounding may bias our results.

Since our method assumes no exposure-mediator interaction, we fitted, for each risk factor, a logistic regression that included the risk factor, mediator, potential confounders and an interaction term involving the risk factor and the mediator. Then we tested for interaction using a standard Wald test.

Table 6 presents the estimated mediation proportions, confidence intervals, p-values, the estimated risk factor effects, and the p-value for risk factor-mediator interaction. MD was a significant mediator (at the 5% significance level) for HBBD, ASM and BMI18, regardless whether the test was based on $\hat{p}$ or $\hat{d}$, although p-values corresponding to the latter test were much smaller. MD was significant as a mediator for AM according to the $\hat{d}$ test but not according to the less powerful $\hat{p}$ test. Confidence intervals were quite wide for the mediation proportions for ASM and BMI18. This may be due to the moderate sample size, and the relatively small effect.

There was no evidence for risk factor-mediator interaction for all but one risk factor studied here. For HT, the test for the interaction term was significant. However, the point estimate for the proportion of the effect of HT mediated through MD is very close to zero. Thus, the fact that this assumption is violated is unlikely to be of substantive importance. In the supplementary materials of [46], it was reported that when taking the interaction into account using the method of [30], the mediation proportion remained very small, although positive.

The results suggest that, if the non-confounding assumptions needed for causal interpretation of observed associations are met, MD mediates the effect of at least some pre-menopausal breast cancer risk factors, with evidence for a large mediation proportion for BMI18 and ASM and some mediation of HBBD, but not for the other risk factors.

7 Discussion

In this paper, we have provided methodology for estimation and inference for the mediation proportion in GLMs and the Cox model using the difference method. Our methodology for GLMs uses a data duplication algorithm with GEE and allows for the consistent estimation of the covariance of the estimates.

Strictly speaking, the validity of the difference method relies on the assumption that the marginal model, the one that does not include the mediator, and the conditional model, the one that does, hold

Table 6: Mediation analysis for pre-menopausal breast cancer incidence with mammographic density as the mediator in the NHS and NHSII studies (N= 559 cases and 1727 controls). $\hat{RR}_{total} = \exp(\hat{\beta}^{\star})$, $\hat{RR}_{direct} = \exp(\hat{\beta})$.

Risk factor	p–inter*	$\hat{\beta}^{\star}$ ($\hat{RR}_{total}$)	p-value	$\hat{\beta}$ ($\hat{RR}_{direct}$)	$\hat{p}$	95% CI	p-value, $\hat{p}$ test	p-value, $\hat{d}$ test
Personal history of benign breast disease	0.95	0.35(1.42)	< 0.001	0.25(1.28)	0.30	0.10–0.51	0.004	$< 10^{-6}$
Family history of breast cancer	0.59	0.42(1.52)	0.01	0.42(1.52)	0.004	−0.10–0.11	0.94	0.94
Adolescent somatotype† Per 3 unit increase	0.20	−0.34(0.72)	0.02	−0.12(0.88)	0.63	0.05– 1.20	0.03	$< 10^{-7}$
BMI at age 18† Per 5 unit increase	0.20	−0.23(0.79)	0.02	−0.05(0.95)	0.78	0.06– 1.50	0.03	$< 10^{-7}$
Age at first birth‡ Per 5 year increase	0.17	0.15(1.17)	0.03	0.15(1.16)	0.03	−0.09–0.15	0.31	0.30
Age at menarche Per 2 year increase	0.22	−0.16(0.86)	0.03	−0.18(0.84)	−0.16	−0.36–0.04	0.12	0.04
Height Per 3 inch increase	0.03	0.13(1.14)	0.03	0.14(1.14)	−0.01	−0.14–0.11	0.82	0.82

Adjusted for age, fasting status, blood collection time of the day, mammography batch (NHS batch 1, NHS batch 2 or NHSII), current and at age 18 BMI, adolescent somatotype, history of BBD, parity, age at first birth, and age at menarche

$\star$ P-value of the test for interaction between percent MD and each risk factor

† Not adjusted for adolescent somatotype, BMI, current or at age 18

‡ Among parous women only (478 cases, 1499 controls)

simultaneously. However, we demonstrate in this paper that *g*-linkability with respect to the mean functions ensures that the point estimator for the mediation proportion is consistent under standard assumptions for the identity and log link functions and under a rare outcome assumption for the logistic link function and Cox model. The rare outcome assumption is fulfilled in most chronic disease incidence studies in epidemiology, including the one motivating the present work. When the outcome is not rare, one may fit the log-binomial model instead, as noted in [38], which may be preferable anyway, as the odds ratio is typically not the parameter of interest [47]. Furthermore, the estimator is asymptotically normally distributed with a variance that can be consistently estimated using a robust sandwich estimator easily obtained by applying a data duplicated GEE. In some scenarios, *g*-linkability fails to hold, even approximately. A direction for future research is alternative definitions and estimation procedures that are based on nonparametric projections instead of exact generalized linear models [48].

Despite its popularity, the difference method for estimating the mediation proportion has been criticized due to what appeared to be undesirable finite samples properties [4, 40]. However, when considering binary outcomes, the covariance (or correlation) between estimates from the marginal and conditional models was typically estimated using approximations from the linear model [1]. We have now developed methodology for a valid covariance estimator and showed that testing for mediation using a test based on the difference yields a valid statistical test, even in finite samples.

An alternative to the difference method is the product method. In terms of finite sample properties, we have shown the two methods are comparable under *g*-linkability. A major advantage of using the difference method is that variance can be directly estimated by standard software, without relying on approximations or on the bootstrap, which are typically used for inference when working with the product method. Thus, we have provided a valid, simple alternative and provided software in **SAS** and **R**.

The causal structure and the underlying confounding assumptions are important to consider when our methods are used in applications. Confounding may occur due to exposure-mediator confounders, exposure-outcome confounders or mediator-outcome confounders. We refer the readers to [19], and references therein, for relevant discussions on assumptions needed and analysis conducted in order to avoid, or at least minimize, potential bias due to confounding when conducting mediation analysis. The difference method does not allow for mediator-exposure interaction, and alternative methods to allow for this interaction were previously developed [34, 38, 49].

In practice, mediation analysis is often conducted for well-established exposures or risk factors, or when the total effect is significant. As suggested by our simulation results, when the total effect was small, mediation analysis was less likely to provide adequate results. On the other hand, an analysis that only considers significant total effects should take into account that it was performed conditionally on the results of a first stage analysis. The properties of such conditional inference can be considered in future research.

In our implementation of the GEE methodology, we propose to use the independence working correlation matrix, which has the nice property of providing identical coefficient estimates when fitting the two models separately and when using the data duplication algorithm, fitting them together. Under other working correlation matrices, this property does not hold anymore, but efficiency may be gained.

In conclusion, the general framework for mediation analysis in GLMs developed in this paper along with the methodology established, will allow researchers to investigate mediation under various outcome scenarios and to quantify results based on rigorously derived and empirically studied estimators and hypothesis tests.

A Appendix

A.1 Software

The **SAS** macro `%mediate` implements the data duplication algorithm and reports point and interval estimates for the mediation proportion and the results for the mediation test using the difference method. It is available

on the last author's website `http://www.hsph.harvard.edu/donna-spiegelman/software/mediate`. The **SAS** macro supports GLMs and the Cox model. We also developed an **R** package named `GEEmediate` which is available on **CRAN**. The current version of the **R** package can be used for **GLMs**.

References

1. Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials or observational studies. Am. J. Epidemiol. 1992;136:1148–1159.
2. MacKinnon DP. Introduction to statistical mediation analysis Routledge, 2008.
3. MacKinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. Eval. Rev. 1993;17:144–158.
4. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol. Meth. 2002;7:83.
5. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Personality Social Psychol. 1986;51:1173.
6. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Stat. Sci. 2010b:51–71.
7. Pearl J. Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 2001:411–420.
8. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992:143–155.
9. Huang B, Sivaganesan S, Succop P, Goodman E. Statistical assessment of mediational effects for logistic mediational models. Stat. Med. 2004;23:2713–2728.
10. Wang W., Albert JM. Estimation of mediation effects for zero-inflated regression models. Stat. Med. 2012;31:3118–3132.
11. Huang Y-T, Pan W-C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 2015.
12. Alwin DF, Hauser RM. The decomposition of effects in path analysis. Am. Sociological Rev. 1975:37–47.
13. Judd CM. Kenny DA. Process analysis estimating mediation in treatment evaluations. Eval. Rev. 1981;5:602–619.
14. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat. Med. 1992;11:167–178.
15. Lin D, Fleming T, De Gruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. Stat. Med. 1997;16:1515–1527.
16. Parast L, McDermott MM, Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. Stat Med. 2015.
17. Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. Soc. Method. 1982;13:290–312.
18. Goodman LA. On the exact variance of products. J. Am. Stat. Assoc. 1960;55:708–713.
19. VanderWeele T. Explanation in causal inference: methods for mediation and interaction. Oxford University Press, 2015.
20. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol. Meth. 2010a;15:309.
21. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. 2014.
22. Carney, RM, Howells WB, Blumenthal JA, Freedland KE, Stein PK, Berkman LF, Watkins LL, Czajkowski SM, Steinmeyer B, Hayano J, et al. Heart rate turbulence, depression, and survival after acute myocardial infarction. Psychosomatic Med. 2007;69:4–9.
23. Lyall K, Ashwood P, Van de Water J, Hertz-Picciotto I. Maternal immune-mediated conditions, autism spectrum disorders, and developmental delay. J. Autism Dev. Disorders 2014;44, 1546–1555.
24. Reisner SL, Greytak EA, Parsons JT, Ybarra ML. Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. J. Sex Res. 2015;52:243–256.
25. Roberts AL, Rosario M, Corliss HL, Koenen KC, Austin SB. Childhood gender nonconformity: A risk indicator for childhood abuse and posttraumatic stress in youth. Pediatrics 2012;129:410–417.
26. Huber PJ. Robust statistics Springer, 2011.
27. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986:13–22.
28. Pearl J. Causal inference in statistics: An overview. Stat. Surv. 2009;3:96–146.
29. Pearl J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prev. Sci. 2012;13:426–436.
30. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am. J. Epidemiol. 2010;172:1339–1348.
31. Valeri, L., X Lin, VanderWeele TJ. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. Stat. Med. 2014;33:4875–4890.
32. Tchetgen Tchetgen EJ. Inverse odds ratio-weighted estimation for causal mediation analysis. Stat. Med. 2013;32:4567–4580.
33. Lange T, Hansen JV. Direct and indirect effects in a survival context. Epidemiology 2011;22:575–581.

34. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. Am. J. Epidemiol. 2012;176:190–195.
35. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. Int. J. Biostat. 2011;7:1–38.
36. VanderWeele TJ. Causal mediation analysis with survival data. Epidemiology (Cambridge, Mass.) 2011;22:582.
37. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. Stat. Meth. Med. Res. 2004;13:309–323.
38. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. Psychol. Methods 2013;18:137.
39. Jiang Z, VanderWeele TJ. When is the difference method conservative for assessing mediation? Am. J. Epidemiol. 2015;182:105–8.
40. Freedman LS. Confidence intervals and statistical power of the "validation" ratio for surrogate or intermediate endpoints. J. Stat. Plann. Inference 2001;96:143–153.
41. MacKinnon DP, Warsi G, Dwyer JH. A simulation study of mediated effect measures. Multivariate Behav. Res. 1995;30, 41–62.
42. Tein J-Y, MacKinnon DP. Estimating mediated effects with survival data. In: New developments in psychometrics. Springer, 2003:405–412.
43. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. Am. J. Nursing 1978;78:1039–1040.
44. Wolf AM, Hunter DJ, Colditz GA, Manson JE, Stampfer MJ, Corsano KA, Rosner B, Kriska A, Willett WC. Reproducibility and validity of a self-administered physical activity questionnaire. Int. J. Epidemiol. 1994;23:991–999.
45. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol. Biomarkers Prev. 2006;15:1159–1169.
46. Rice MS, Bertrand KA, VanderWeele TJ, Rosner BA, Liao X, Adami H-O, Tamimi RM. Mammographic density and breast cancer risk: a mediation analysis. Breast Cancer Res. 2016;18:94.
47. Spiegelman D, Hertzmark E. Easy sas calculations for risk or prevalence ratios and differences. Am. J. Epidemiol. 2005;162:199–200.
48. Stone CJ. The dimensionality reduction principle for generalized additive models. Ann. Stat. 1986:590–606.
49. Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: An r package for flexible mediation analysis using natural effect models. J. Stat. Softw. 2017;76:1–46.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2017-0006).