Harvard University
Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 204

Estimation and Inference for the Mediation Proportion

Daniel Nevo* Xiaomei Liao†
Donna Spiegelman‡

*Harvard School of Public Health, nhdne@channing.harvard.edu
†Harvard School of Public Health, stxia@channing.harvard.edu
‡Harvard School of Public Health, stdls@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper204

Estimation and Inference for the Mediation Proportion

Daniel Nevo, Xiaomei Liao, and Donna Spiegelman

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. We developed estimation and inference methods that reflect the fact the mediation proportion is bounded between zero and one. In particular, we developed statistical testing procedures for the existence of mediation that honors these bounds, and compare the empirical behavior of crude and logit based confidence intervals. 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 at the last author's website.

Estimation and inference for the mediation proportion

Daniel Nevo, Xiaomei Liao and Donna Spiegelman

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. We developed estimation and inference methods that reflect the fact the mediation proportion is bounded between zero and one. In particular, we developed statistical testing procedures for the existence of mediation that honors these bounds, and compare the empirical behavior of crude and logit based confidence intervals. 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 at the last author's website.

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 [21, 9, 18, 20]. Baron and Kenny [2] described a sequence of hypothesis tests to asses 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 have been changed if the exposure value were increased by one unit [30, 24]. 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. Examples include logistic regression [10], zero-inflated regression models [37] and high-dimensional mediators in linear regression with normal errors [11].

One way to estimate mediation is through the "product method" [2]. Another widely used method for assessing mediation is the "difference method" [2, 1, 14]. 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 [8]. A time-to-event final outcome with surrogate biomarkers was also considered [16]. They authors of [16] 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.,[5, 23]).

While the mediation proportion is a popular measure in mediation analysis [4, 28, 29, 17, 25], statistical inference for this parameter is not sufficiently developed. Early important contributions include [16] for Cox regression and [8] for logistic regression. 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. We consider aspects of practical use and compare alternative methods for construction of confidence intervals and statistical tests. Our investigation of these aspects is expanded beyond GLMs to inference about the mediation proportion for Cox model [6].

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, comparing the finite sample properties between possible procedures. In Section 6, we illustrate

the use of the methodology developed in studying mediation of the effect of risk factors for premenopausal breast cancer incidence by mammographic density in the Nurses' Health Studies NHSI [3] and NHSII [38]. In Section 7 we discuss results and related issues. We describe the software we make 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 $\boldsymbol{Z}_i = (X_i, M_i, \boldsymbol{W}_i)$ where X_i, M_i and $\boldsymbol{W}_i$ are an exposure of interest, a mediator and a vector of confounders, respectively. We assume the conditional mean function for the outcome is $E(Y_i|\boldsymbol{Z}_i) = g^{-1}(\boldsymbol{Z}_i^T\boldsymbol{\beta})$ with g being the *link function* and where $\boldsymbol{\beta}$, an unknown parameter vector, is composed of the appropriate components β_X,β_M and $\boldsymbol{\beta_W}$. A consistent estimator, $\hat{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$ is obtained as the solution to the estimating equations

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

where $\boldsymbol{D}_i = \partial E(Y_i|\boldsymbol{Z}_i)/\partial\boldsymbol{\beta}$ and v_i is a working variance of y_i. By GEE theory, the variance of $\hat{\boldsymbol{\beta}}$ can be consistently estimated by the robust sandwich estimator [15, 12].

Traditionally, mediation analysis considers a single mediator. However, methods for addressing multiple mediators have been developed [35]. For simplicity of presentation, we consider in this paper the case of a single mediator M. First, consider the following conditional and marginal mean models for Y, with respect to M

$$E(Y|X, M, \boldsymbol{W}) = g^{-1}(\beta_0 + \beta_1 X + \beta_2 M + \boldsymbol{\beta}_3^T\boldsymbol{W}) \tag{2}$$

$$E(Y|X, \boldsymbol{W}) = g^{-1}(\beta_0^\star + \beta_1^\star X + \boldsymbol{\beta}_3^{\star T}\boldsymbol{W}). \tag{3}$$

Let $\mathcal{B} = (\beta_0, \beta_1, \beta_2, \boldsymbol{\beta}_3)$ and $\mathcal{B}^\star = (\beta_0^\star, \beta_1^\star, \boldsymbol{\beta}_3^\star)$ be the vectors of conditional and marginal regression model parameters, and denote $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^\star$ for their estimators obtained by solving equation (1) under models (2) and (3), separately, respectively. When the two models (2) and (3) both hold simultaneously, we say we have *g-linkability*.

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. Under these assumptions, and when models (2) and (3) hold, the TE equals $\beta_1^\star$ and the NIE equals $\beta_1^\star - \beta_1$, for the identity and log link functions [19, 34], and if, in addition, the outcome is rare, this is also true for the logit link function [36]. Therefore, the mediation proportion, p, which is the ratio between the NIE and the TE, equals to

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

A necessary condition for M to be interpreted as a mediator is that $p \in (0, 1]$. 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 $p \notin [0, 1]$, it is either that the NDE and NIE are in opposite directions or M is not a mediator at all, but

a confounder. In this paper, we only consider the more common and more interpretable case, where the NDE and the NIE are in the same direction. 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 $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^\star$. 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. Lin et al. [16] considered this question for the Cox model in the context of the PTE. First, as in [16], we define the following two models for the hazard function at time t, $h(t)$, conditionally and marginally, with respect to M

$$\begin{aligned} h(t|X, M, \boldsymbol{W}) &= \lambda_0(t) \exp(\beta_0 + \beta_1 X + \beta_2 M + \boldsymbol{\beta}_3^T \boldsymbol{W}) \\ h(t|X, \boldsymbol{W}) &= \lambda_0^\star(t) \exp(\beta_0^\star + \beta_1^\star X + \boldsymbol{\beta}_3^{\star T} \boldsymbol{W}), \end{aligned} \tag{4}$$

where X, M and $\boldsymbol{W}$ are allowed to be time dependent and $\lambda_0(t)$ and $\lambda_0^\star(t)$ are baseline hazard functions. The authors of [16] have shown that these two models cannot hold at same time. However, they claimed that if either $\boldsymbol{\beta}_3^\star$ or $\Lambda_0^\star(t) = \int_0^t \lambda_0^\star(s)ds$ are small, then model (4) 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 (4) 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 [27], 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 [33]), then one may expect $\overline{p}$ to be close to p, as discussed in [16] 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 $\boldsymbol{W}$ 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 [13]. 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

$$E(Y|X, M, \boldsymbol{W}) = \beta_0 + \beta_1 X + \beta_2 M + \boldsymbol{\beta}_3^T \boldsymbol{W}$$
$$E(Y|X, \boldsymbol{W}) = \beta_0^\star + \beta_1^\star X + \boldsymbol{\beta}_3^{\star T} \boldsymbol{W}.$$

We now show that g-linkability holds whenever $E(M|X, \boldsymbol{W})$ is a linear function of X and $\boldsymbol{W}$. To see that, let $E(M|X, \boldsymbol{W}) = a + b_1 X + \boldsymbol{b}_3^T W$, for some a, b_1 and $\boldsymbol{b}_3$. Then,

$$E(Y|X, \boldsymbol{W}) = E(E(Y|X, M, \boldsymbol{W})|X, \boldsymbol{W}) = \beta_0 + \beta_1 X + \beta_2 E(M|X, \boldsymbol{W}) + \boldsymbol{\beta}_3^T \boldsymbol{W} = \beta_0^\star + \beta_1^\star X + \boldsymbol{\beta}_3^{\star T} \boldsymbol{W}$$

where $\beta_0^\star = \beta_0 + a$, $\beta_1^\star = \beta_1 + b_1$ and $\boldsymbol{\beta}_3^\star = \boldsymbol{\beta}_3 + \boldsymbol{b}_3$.

3.2 Log link function

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

$$E(Y|X, M, \boldsymbol{W}) = \exp(\beta_0 + \beta_1 X + \beta_2 M + \boldsymbol{\beta}_3^T \boldsymbol{W})$$
$$E(Y|X, \boldsymbol{W}) = \exp(\beta_0^\star + \beta_1^\star X + \boldsymbol{\beta}_3^{\star T} \boldsymbol{W})$$

and we have

$$E(Y|X, \boldsymbol{W}) = E(E(Y|X, M, \boldsymbol{W})|X, \boldsymbol{W}) = \exp(\beta_0 + \beta_1 X + \boldsymbol{\beta}_3^T \boldsymbol{W}) \times E[\exp(\beta_2 M)|X, \boldsymbol{W}].$$

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

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 [27, 33, 13]. The logit link function, defined as $\text{logit}(p) = \log(p/(1-p))$, 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 [36]. 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 $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^\star$ defined in Section 2. By the aforementioned GEE theory together with the general mapping theorem, this estimator is consistent.

Confidence intervals for p were constructed in the past using either Fieller's theorem or the delta method [8, 16, 7]. Here, we consider the latter. As written in [16], by the delta method $\hat{p}$

has an asymptotic normal distribution with variance equals to

$$\sigma^2_{\hat{p}} = \frac{\sigma^2_{\hat{\beta}_1}}{(\beta^\star_1)^2} + \frac{\beta^2_1 \sigma^2_{\hat{\beta}^\star_1}}{(\beta^\star_1)^4} - 2\frac{\beta_1 \sigma_{\hat{\beta}_1, \hat{\beta}^\star_1}}{(\beta^\star_1)^3}, \tag{5}$$

where

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1), \quad \sigma^2_{\hat{\beta}^\star_1} = Var(\hat{\beta}^\star_1) \quad \text{and} \quad \sigma_{\hat{\beta}_1, \hat{\beta}^\star_1} = Cov(\hat{\beta}_1, \hat{\beta}^\star_1).$$

While $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}^\star_1}$ can be consistently estimated using the robust sandwich estimator [12] for each of the models (2) and (3) separately, it is not obvious how to estimate $\sigma_{\hat{\beta}_1, \hat{\beta}^\star_1}$, the covariance of estimators obtained from two separate models. In Section 4.2, we propose a data duplication algorithm to estimate this quantity.

Assume now we have estimates $\hat{\sigma}^2_{\hat{\beta}_1}, \hat{\sigma}^2_{\hat{\beta}^\star_1}$ and $\hat{\sigma}_{\hat{\beta}_1, \hat{\beta}^\star_1}$ for $\sigma^2_{\hat{\beta}_1}, \sigma^2_{\hat{\beta}^\star_1}$ and $\sigma_{\hat{\beta}_1, \hat{\beta}^\star_1}$, respectively. These estimates are plugged in (5) in order to get an estimate $\hat{\sigma}^2_p$ and a $(1-\alpha)$ level confidence interval for p may be obtained as

$$\hat{p} \pm z_{1-\alpha/2} \hat{\sigma}_{\hat{p}} \tag{6}$$

with $z_{1-\alpha/2}$ being the appropriate quantile of the normal distribution. While this confidence interval is asymptotically valid, in finite samples it may include negative values or values larger than one. Since such values are outside the parameter space for p if M is indeed a mediator, values outside the parameter space should be excluded. One option is to trim the resulting confidence interval so it would be contained in $[0, 1]$. Alternatively, a logit-based confidence interval can be constructed, using again the delta method, and then back-transformed the resulting confidence interval to get a confidence interval which is, by definition, contained in $[0, 1]$. More formally, let $\psi = \text{logit}(p) = \log(\frac{p}{1-p})$ and let $\hat{\psi} = \text{logit}(\hat{p})$. By the delta method, $\hat{\psi}$ is consistent and asymptotically normally distributed with variance

$$\sigma^2_{\hat{\psi}} = \frac{1}{p^2(1-p)^2} \sigma^2_{\hat{p}},$$

which can be estimated by plugging in $\hat{p}$ and $\hat{\sigma}^2_p$ in the above expression. Then, a $(1-\alpha)$ level confidence interval for the mediation proportion p is obtained as

$$\left[\frac{\exp\left(\hat{\psi} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\psi}}\right)}{1 + \exp\left(\hat{\psi} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\psi}}\right)}, \frac{\exp\left(\hat{\psi} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\psi}}\right)}{1 + \exp\left(\hat{\psi} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\psi}}\right)} \right] \tag{7}$$

4.1 Hypothesis testing

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 [8, 16, 7]. 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. $H_1 : p > 0$. Now, recall that $p \in [0, 1]$. Let $Z_p = \sigma_p^{-1}\hat{p}$ be the scaled unconstrained estimate. By the delta method, the distribution of Z_p under the null converges to a standard normal distribution. Let $Z_p^+ = \max(0, Z_p)$ be that scaled estimate, obtained by replacing the unconstrained $\hat{p}$ by zero if $\hat{p}$ is negative. By the general

mapping theorem, we then have

$$\lim_{n\to\infty} P(Z_p^+ \geq c) \to \begin{cases} 1 & c \leq 0, \\ 1-\Phi(c) & c > 0 \end{cases}$$

where Φ is the cumulative distribution function of standard normal distribution and the p-value for testing $H_0 : p = 0$ vs. $H_1 : p > 0$ equals to one if $\hat{\sigma}_p^{-1}\hat{p} < 0$ and to $1 - \Phi(\hat{\sigma}_p^{-1}\hat{p})$ if $\hat{p} > 0$. However, if the unconstrained $\hat{p}$ is larger than one, then this is misleading. Indeed p maybe larger than zero in this case, but p should not be interpreted as a mediation proportion, as noted in [16]. In this case, M is not a mediator, but a confounder. This demonstrates the point that the described test should not be used without first considering the unconstrained value of $\hat{p}$, and making sure it is within the parameter space $[0, 1]$.

Consider the distribution of $(Z_p^+)^2 = [\max(0, Z_p)]^2$. This statistic equals to zero if $Z_p < 0$, which occurs with probability of 0.5 since Z_p is a standard normal variable. Therefore, for any nonnegative value c^+ we have

$$P[(Z_p^+)^2 < c^+] = P[Z_p \leq 0] + P[0 < Z_p < \sqrt{c^+}] = 0.5 + 0.5P[Z_p^2 < c^+] = 0.5 + 0.5P(\chi^2_{(1)} < c^+)$$

since $P[0 < Z_p < \sqrt{c^+}] = 0.5P[-\sqrt{c^+} < Z_p < \sqrt{c^+}]$, where $\chi^2_{(k)}$ is a χ^2 variable with k degrees of freedom. Thus, the asymptotic distribution of $(Z_p^+)^2$ is a mixture of $\chi^2_{(0)}$ and $\chi^2_{(1)}$ random variables, with mixture probability of 0.5, similar to what was previously shown [31].

An alternative test statistic is based upon a test for the difference between the effect estimates in the marginal and the 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

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

Assume that $\hat{\beta}_1^\star > 0$. Similar to Z_p^+, we define $Z_d^+ = \max(0, Z_d)$ and then the p-value of the difference test is calculated as $1 - \Phi(\hat{\sigma}_d^{-1}\hat{d})$, if $d > 0$ and 1 otherwise. It is readily observed that $\hat{d} \in (0, \hat{\beta}_1^\star]$ if and only if $\hat{p} \in (0, 1]$. Thus, one should treat a crude result of $\hat{d} \notin (0, \hat{\beta}_1^\star]$ as he or she would have treated the situation $\hat{p} \notin (0, 1]$. Finally, if the coefficient in model (3) corresponding to the effect of X is negative, then the p-value for the difference test is $\Phi(\hat{\sigma}_d^{-1}\hat{d})$ if $\hat{d} < 0$ and zero otherwise.

4.2 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 $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^\star$ 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 [16] 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, 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 [15],

$$E(Y_{ij}|X_i, X_i^\star, M_i, \boldsymbol{W}_i, \boldsymbol{W}_i^\star) = g^{-1}(\beta_0 I\{j=1\}+\beta_1 X_i+\beta_2 M_i+\boldsymbol{\beta}_3^T \boldsymbol{W}_i+\beta_0^\star I\{j=2\}+\beta_1^\star X_i^\star+\boldsymbol{\beta}_3^{\star T}\boldsymbol{W}_i^\star), \tag{8}$$

where $j = 1, 2$ are the rows created from duplicating each observation and are treated as repeated measures. Model (8) implies that we can write $E(Y_{i1}|X_i, X_i^\star, M_i, \boldsymbol{W}_i, \boldsymbol{W}_i^\star) = E(Y_{i1}|X_i, M_i, \boldsymbol{W}_i)$ and $E(Y_{i2}|X_i, X_i^\star, M_i, \boldsymbol{W}_i, \boldsymbol{W}_i^\star) = E(Y_{i2}|X_i^\star, \boldsymbol{W}_i^\star)$. Let $\boldsymbol{R}$ be a 2×2 working correlation matrix and denote $\boldsymbol{B}_i = diag(v_{i1}, v_{i2})$, where $v_{ij} = Var(Y_{ij})$. Let also $\boldsymbol{V}_i = \boldsymbol{B}_i^{1/2}\boldsymbol{R}\boldsymbol{B}_i^{1/2}$ be a 2×2 working variance for the vector (Y_{i1}, Y_{i2}). Here, the GEE are defined as

$$U_{GEE}(\mathcal{B}) = \sum_{i=1}^{n}(\mathcal{D}_i, \mathcal{D}_i^\star)\boldsymbol{V}_i^{-1}\begin{pmatrix} y_{i1} - E(Y_{i1}|X_i, M_i, \boldsymbol{W}_i) \\ y_{i2} - E(Y_{i2}|X_i^\star, \boldsymbol{W}_i^\star)\end{pmatrix}, \tag{9}$$

where $\mathcal{D}_i = \partial E(Y_{i1}|X_i, M_i, \boldsymbol{W}_i)/\partial\mathcal{B}$ and $\mathcal{D}_i^\star = \partial E(Y_{i2}|X_i^\star, \boldsymbol{W}_i^\star)/\partial\mathcal{B}^\star$ are two column vectors. If $\boldsymbol{R}$ is taken to be the identity matrix, then $\boldsymbol{V}_i = \boldsymbol{B}_i$ and (9) simplifies to the following estimating equations

$$U_{IEE}(\mathcal{B}) = \begin{pmatrix} U_{IEE}^{(1)}(\boldsymbol{\beta}) \\ U_{IEE}^{(2)}(\boldsymbol{\beta}^\star)\end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \mathcal{D}_i v_{i1}^{-1}[y_i - E(Y_{i1}|X_i, M_i, \boldsymbol{W}_i)] \\ \sum_{i=1}^n \mathcal{D}_i^\star v_{i2}^{-1}[y_i - E(Y_{i2}|X_i^\star, \boldsymbol{W}_i^\star)]\end{pmatrix} = 0. \tag{10}$$

Then, the estimating equations given by (10) are identical to the estimating equations for fitting models (2) and (3) separately, because $\boldsymbol{D}_i$ and v_i and $\boldsymbol{Z}_i$ in equation (1) are equal to $\mathcal{D}_i, v_{i1}$ and $(X_i, M_i, \boldsymbol{W}_i)$, respectively, under model (2), and they are equal to $\mathcal{D}_i^\star, v_{i2}$ and $(X_i^\star, \boldsymbol{W}_i^\star)$, 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 $\hat{\mathcal{B}}$, 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

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 intervals and the type I error and the power of the statistical tests described in Section 4. For the generalized linear models, we used the GEE data duplication method as described in the previous section. For the Cox model, estimates were calculated using the data duplication method suggested by Lin et al. [16].

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 asses 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 4.2 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

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

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

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

where $\text{expit}(u) = \exp(u)/(1+\exp(u))$. 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. $p = 0.1, 0.2, ..., 0.8$; $\rho = p, p+0.1, ..., 0.8$, 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.

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 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.

Table 2 presents relative bias of $\hat{p}$ and coverage rates of the trimmed version of the 95% confidence intervals given by (6). The mediation proportion estimates exhibited unstable performance whenever the sample size was low and the total effect of X, $\beta_1^\star$, was small, regardless of the magnitude of the mediation proportion. However, these problems disappeared when we considered a larger sample size or a larger total effect of X. The results for the logit link function and the Cox model in Table 2 are presented for the rare outcome case.

From estimation we move to hypothesis testing. The two test statistics compared were described in Section 4.1, where the variance estimators used in the test statistics were obtained by the data duplication algorithm described in Section 4.2. As previously discussed, both tests used were one-sided since the mediation proportion is nonnegative. Since p must be in $[0, 1]$ in both cases, if $\hat{p}$ was outside this interval we did not conducted the tests. The last column in Table 2 gives the mean proportion of such simulations, where the mean is taken over the other columns. The proportion of simulations with $\hat{p} \notin [0, 1]$ was larger when the total effect of X and the sample size were small. 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 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 intervals proposed in Section 4. We considered an untransformed confidence interval (6) as well as a logit transformation based confidence interval (7). Table 4 compares these two options in terms of coverage rate and confidence interval width. When the unconstrained point estimate was negative or larger than one, i.e., $\hat{p} \notin [0, 1]$, a confidence interval was not constructed. Crude confidence intervals were trimmed to be contained in $[0, 1]$, while the logit transformation-based confidence intervals are contained in $[0, 1]$ by definition. Coverage rates were generally adequate. When ρ was much

larger than p, neither of the methods produced confidence intervals with nominal coverage, especially when the sample size was small. Comparing between the trimmed untransformed and the transformed-based confidence intervals, the latter did not offer any clear advantage in terms of performance. For small sample size, the transformation-based confidence interval tended to be wider, especially for $p = 0.1$, without any substantial gain in terms of coverage rate. For a larger sample size, the two confidence intervals were comparable in their performance.

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 C.

6 Illustrative example

We illustrate the use of our methodology in the analysis of breast cancer data from the Nurses Health's Studies (NHS and NHSII) [3, 38]. It was previously found that high mammographic density (MD) is a risk factor for breast cancer [22]. 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. Detailed description of this study is given in [26]. 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, and as shown in the previous section, g-linkability should hold. The mediator is percent MD. We conducted mediation analysis for all breast cancer 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). Results were adjusted for 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, where mediation was assessed separately for a number of these variables, where most of the others were treated as confounders.

Table 5 presents the estimated mediation proportions, confidence intervals and p-values, along with the estimated risk factor effects. Of note is that MD is significant as a mediator 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. Confidence intervals were quite wide for ASM and BMI18. This may be due to the moderate sample size, and the relatively small effect.

7 Discussion

In this paper, we have provided methodology for estimation and inference for the mediation proportion in generalized linear models 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 simultaneously. To address this concern, researchers have suggested more complicated methods, e.g, a Bayesian approach [5], and more recently, a nonparamteric method [23]. 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. 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.

Despite its popularity, the difference method for estimating the mediation proportion has been criticized due to what appeared to be undesirable finite samples properties [20, 7]. 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 [9]. 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.

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-outcome confounders or mediator-outcome confounders. We refer the readers to [34], 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 [33].

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.

Our approach assumes that the underlying model is g-linkable, where the same link function

holds for both the conditional model and the marginal model, where the latter is the model without the mediator. We have shown that g-linkability holds for the identity and the log link function when fairly general conditions are met. In addition, g-linkability holds for the logit link function whenever the outcome is rare. When the outcome is not rare, one may fit the log-binomial model instead, as noted in [33], which may be preferable anyway, as the odds ratio is typically not the parameter of interest [32].

In conclusion, the general framework for mediation analysis in generalized linear models 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.

Acknowledgments

This work was supported by National Institutes of Health grant DP1ES025459.

Appendix

One major goal of this paper is to produce statistical tools to be used in practice. 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`. Simulations were conducted using **R** code that can be obtained by request to the first author. Both the **SAS** macro and the **R** code can be used for either GLMs or survival data analysis.

References

[1] Duane F Alwin and Robert M Hauser. The decomposition of effects in path analysis. *American sociological review*, pages 37–47, 1975.

[2] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[3] Charlene F Belanger, Charles H Hennekens, Bernard Rosner, and Frank E Speizer. The nurses' health study. *The American Journal of Nursing*, 78(6):1039–1040, 1978.

[4] Robert M Carney, William B Howells, James A Blumenthal, Kenneth E Freedland, Phyllis K Stein, Lisa F Berkman, Lana L Watkins, Susan M Czajkowski, Brian Steinmeyer, Junichiro Hayano, et al. Heart rate turbulence, depression, and survival after acute myocardial infarction. *Psychosomatic medicine*, 69(1):4–9, 2007.

[5] Mary Kathryn Cowles. Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Statistics in medicine*, 21(6):811–834, 2002.

[6] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[7] Laurence S Freedman. Confidence intervals and statistical power of the validationratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, 96(1):143–153, 2001.

[8] Laurence S Freedman, Barry I Graubard, and Arthur Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2):167–178, 1992.

[9] Laurence S Freedman and Arthur Schatzkin. Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology*, 136(9):1148–1159, 1992.

[10] Bin Huang, Siva Sivaganesan, Paul Succop, and Elizabeth Goodman. Statistical assessment of mediational effects for logistic mediational models. *Statistics in medicine*, 23(17):2713–2728, 2004.

[11] Yen-Tsung Huang and Wen-Chi Pan. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 2015.

[12] Peter J Huber. *Robust statistics*. Springer, 2011.

[13] Zhichao Jiang and Tyler J VanderWeele. When is the difference method conservative for assessing mediation? *American journal of epidemiology*, page kwv059, 2015.

[14] Charles M Judd and David A Kenny. Process analysis estimating mediation in treatment evaluations. *Evaluation review*, 5(5):602–619, 1981.

[15] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22, 1986.

[16] DY Lin, TR Fleming, and V De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, 16(13):1515–1527, 1997.

[17] Kristen Lyall, Paul Ashwood, Judy Van de Water, and Irva Hertz-Picciotto. Maternal immune-mediated conditions, autism spectrum disorders, and developmental delay. *Journal of autism and developmental disorders*, 44(7):1546–1555, 2014.

[18] David P MacKinnon and James H Dwyer. Estimating mediated effects in prevention studies. *Evaluation review*, 17(2):144–158, 1993.

[19] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annual review of psychology*, 58:593, 2007.

[20] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1):83, 2002.

[21] David Peter MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2008.

[22] Valerie A McCormack and Isabel dos Santos Silva. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*, 15(6):1159–1169, 2006.

[23] Layla Parast, Mary M McDermott, and Lu Tian. Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in medicine*, 2015.

[24] Judea Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

[25] Sari L Reisner, Emily A Greytak, Jeffrey T Parsons, and Michele L Ybarra. Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. *The Journal of Sex Research*, 52(3):243–256, 2015.

[26] Megan S. Rice, Kimberly A. Bertrand, Tyler J. VanderWeele, Bernard Rosner, Xiaomei Liao, Hans-Olov Adami, and Rulla M. Tamimi. Mammographic density and breast cancer risk: A mediation analysis. *Under review*, 2016.

[27] John Ritz and Donna Spiegelman. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13(4):309–323, 2004.

[28] Andrea L Roberts, Margaret Rosario, Heather L Corliss, Karestan C Koenen, and S Bryn Austin. Childhood gender nonconformity: A risk indicator for childhood abuse and posttraumatic stress in youth. *Pediatrics*, 129(3):410–417, 2012.

[29] Andrea L Roberts, Margaret Rosario, Heather L Corliss, Karestan C Koenen, and S Bryn Austin. Elevated risk of posttraumatic stress in sexual minority youths: mediation by childhood abuse and gender nonconformity. *American journal of public health*, 102(8):1587–1593, 2012.

[30] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

[31] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[32] Donna Spiegelman and Ellen Hertzmark. Easy sas calculations for risk or prevalence ratios and differences. *American journal of epidemiology*, 162(3):199–200, 2005.

[33] Linda Valeri and Tyler J VanderWeele. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological methods*, 18(2):137, 2013.

[34] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[35] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115, 2013.

[36] Tyler J VanderWeele and Stijn Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172(12):1339–1348, 2010.

[37] Wei Wang and Jeffrey M Albert. Estimation of mediation effects for zero-inflated regression models. *Statistics in medicine*, 31(26):3118–3132, 2012.

[38] Anne M Wolf, David J Hunter, Graham A Colditz, Joann E Manson, Meir J Stampfer, Karen A Corsano, Bernard Rosner, Andrea Kriska, and Walter C Willett. Reproducibility and validity of a self-administered physical activity questionnaire. *International journal of epidemiology*, 23(5):991–999, 1994.

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 (8).

i	j	Intercept	Intercept*	X	$X^\star$	M	$\boldsymbol{W}$	$\boldsymbol{W}^\star$	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

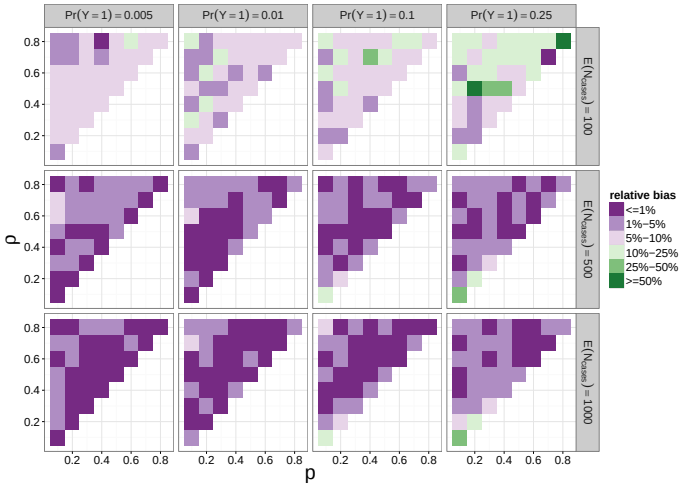

Figure 1: Relative bias of the mediation proportion estimator under the logistic model as a function of the mediation proportion (p), the correlation between the exposure and the mediator (ρ), the number of expected cases ($E(N_{cases})$) and the outcome rate ($P(Y = 1)$). The value of $\beta_1^\star$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

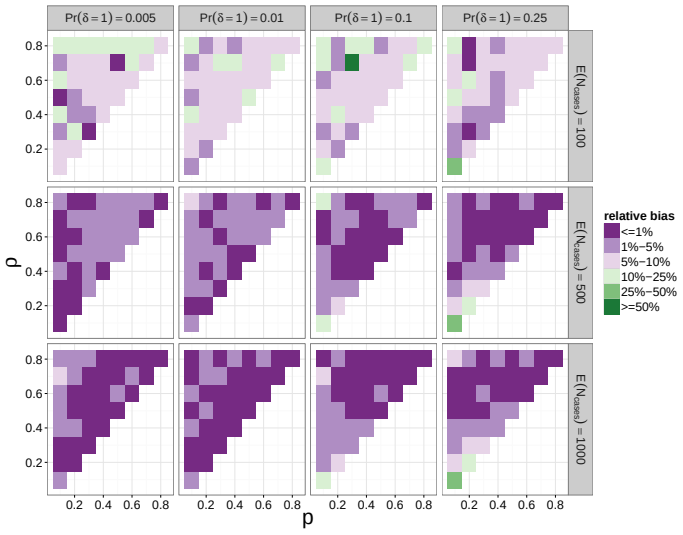

Figure 2: Relative bias of the mediation proportion estimator under the Cox model as a function of the mediation proportion (p), the correlation between the exposure and the mediator (ρ), the number of expected cases ($E(N_{cases})$) and the event rate ($P(\delta = 1)$). The value of $\beta_1^\star$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

Table 2: Relative bias in percentage of the mediation proportion estimator under the identity and logit link functions and the Cox model. Coverage rates of 95% trimmed untransformed confidence intervals are displayed in brackets. N_{out} is the mean proportion of simulations with $\hat{p} \notin [0, 1]$, where the mean is taken over the rest of the columns.

Identity link function ($Y \sim Normal$)

		$n = 1000$			$n = 5000$			
		$\beta_1^\star$			$\beta_1^\star$			N_{out}
		0.1	0.3	0.5	0.1	0.3	0.5	
$p = 0.1$	$\rho = 0.1$	15.3 (0.92)	-0.1 (0.96)	2.3 (0.96)	3.1 (0.95)	0 (0.96)	-1 (0.95)	0.0
	$\rho = 0.3$	11.2 (0.83)	0.3 (0.94)	-0.5 (0.94)	3.6 (0.95)	-0.2 (0.95)	-0.1 (0.94)	2.9
	$\rho = 0.5$	23.9 (0.71)	-5.1 (0.91)	-1.3 (0.94)	1.5 (0.88)	2.6 (0.95)	0 (0.95)	7.9
	$\rho = 0.7$	17.6 (0.6)	0.3 (0.78)	0.8 (0.93)	17 (0.78)	-1.6 (0.96)	-1.1 (0.94)	14.2
$p = 0.3$	$\rho = 0.3$	35.8 (0.92)	1.2 (0.95)	0.3 (0.95)	3 (0.96)	0.4 (0.95)	0.1 (0.94)	0.4
	$\rho = 0.5$	22.3 (0.9)	1.4 (0.95)	0.7 (0.95)	0.9 (0.96)	0.4 (0.96)	-0.5 (0.95)	1.4
	$\rho = 0.7$	34.1 (0.76)	-0.7 (0.94)	-0.8 (0.94)	1.2 (0.96)	1.4 (0.94)	0 (0.94)	4.4
$p = 0.5$	$\rho = 0.5$	15.4 (0.86)	1.5 (0.95)	0.3 (0.95)	2.1 (0.96)	0.1 (0.95)	0.2 (0.95)	1.6
	$\rho = 0.7$	14.8 (0.81)	0.5 (0.95)	0.4 (0.95)	2.5 (0.95)	-0.2 (0.94)	0.2 (0.95)	3.0

Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.005$

		$E(N_{cases}) = 500$			$E(N_{cases}) = 1000$			
		$\beta_1^\star$			$\beta_1^\star$			N_{out}
		log(1.25)	log(1.5)	log(2)	log(1.25)	log(1.5)	log(2)	
$p = 0.1$	$\rho = 0.1$	4.7 (0.94)	-0.4 (0.95)	-0.1 (0.95)	-3 (0.92)	-5.1 (0.9)	-5.3 (0.86)	0.0
	$\rho = 0.3$	5.1 (0.9)	2.9 (0.95)	1.1 (0.96)	2.7 (0.96)	0.5 (0.95)	0.3 (0.95)	0.9
	$\rho = 0.5$	7.6 (0.77)	3.1 (0.93)	-1.7 (0.96)	7.3 (0.87)	-1.4 (0.95)	0.8 (0.96)	5.6
	$\rho = 0.7$	2.6 (0.62)	6.8 (0.82)	-1.1 (0.9)	3.3 (0.74)	0.3 (0.93)	-0.6 (0.93)	14.3
$p = 0.3$	$\rho = 0.3$	7.5 (0.94)	0.8 (0.94)	0.5 (0.95)	0.5 (0.94)	-1.2 (0.93)	-1.0 (0.94)	0.0
	$\rho = 0.5$	8.5 (0.95)	0.9 (0.95)	0.6 (0.95)	1.6 (0.96)	0.1 (0.96)	0.4 (0.94)	0.1
	$\rho = 0.7$	5.9 (0.88)	2.2 (0.95)	-0.4 (0.95)	6.5 (0.96)	-0.8 (0.95)	-0.7 (0.95)	1.3
$p = 0.5$	$\rho = 0.5$	9.2 (0.9)	1.1 (0.96)	0.9 (0.94)	1.5 (0.96)	-0.4 (0.95)	-0.1 (0.96)	0.2
	$\rho = 0.7$	9.1 (0.88)	1.2 (0.96)	0.2 (0.94)	3.5 (0.96)	1.0 (0.95)	0.5 (0.96)	1.0

Cox model with $P(\delta = 1) = 0.005$

		$E(N_{cases}) = 500$			$E(N_{cases}) = 1000$			
		$\beta_1^\star$			$\beta_1^\star$			N_{out}
		log(1.25)	log(1.5)	log(2)	log(1.25)	log(1.5)	log(2)	
$p = 0.1$	$\rho = 0.1$	6.9 (0.93)	0.5 (0.94)	-1.1 (0.93)	-5.5 (0.92)	-7.3 (0.89)	-7.4 (0.85)	0.0
	$\rho = 0.3$	4.9 (0.9)	0.2 (0.95)	-0.2 (0.94)	2.8 (0.96)	-1.2 (0.95)	-0.3 (0.96)	1.3
	$\rho = 0.5$	0.7 (0.77)	0.0 (0.93)	2.6 (0.94)	5 (0.86)	1.3 (0.96)	1.5 (0.95)	6.2
	$\rho = 0.7$	4.2 (0.65)	-1.0 (0.81)	5.6 (0.88)	5.5 (0.74)	4.4 (0.92)	-0.3 (0.94)	14.7
$p = 0.3$	$\rho = 0.3$	0.6 (0.93)	1.3 (0.95)	0.7 (0.97)	-0.4 (0.94)	-1.5 (0.94)	-1.4 (0.95)	0.0
	$\rho = 0.5$	5.7 (0.94)	1.3 (0.94)	0.7 (0.95)	2 (0.96)	0.5 (0.94)	-0.6 (0.96)	0.1
	$\rho = 0.7$	7.7 (0.89)	2.1 (0.94)	-0.6 (0.96)	2.4 (0.94)	-0.2 (0.96)	-1.0 (0.95)	1.6
$p = 0.5$	$\rho = 0.5$	6.5 (0.91)	1.2 (0.96)	0.5 (0.96)	2.4 (0.95)	-0.5 (0.94)	-0.4 (0.96)	0.3
	$\rho = 0.7$	0.7 (0.88)	1.3 (0.94)	0.2 (0.95)	2.4 (0.95)	0.6 (0.94)	0.6 (0.95)	0.8

Table 3: Type I error and power for tests for mediation under the identity and logit link functions and the Cox model.

Identity link function ($Y \sim Normal$)

		$n = 1000$					
		$\beta_1^\star = 0.1$		$\beta_1^\star = 0.3$		$\beta_1^\star = 0.5$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.01	0.03	0.02	0.02	0.02	0.02
	$\rho = 0.3$	0.03	0.06	0.04	0.04	0.05	0.05
	$\rho = 0.5$	0.02	0.05	0.04	0.04	0.05	0.05
	$\rho = 0.7$	0.02	0.04	0.04	0.05	0.04	0.04
$p = 0.1$	$\rho = 0.1$	0.55	0.79	0.95	0.94	0.96	0.95
	$\rho = 0.3$	0.16	0.26	0.90	0.90	1.00	1.00
	$\rho = 0.5$	0.08	0.15	0.46	0.47	0.85	0.85
	$\rho = 0.7$	0.04	0.09	0.26	0.26	0.48	0.48
$p = 0.2$	$\rho = 0.3$	0.45	0.63	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.17	0.27	0.93	0.93	1.00	1.00
	$\rho = 0.7$	0.08	0.14	0.60	0.61	0.95	0.95
$p = 0.3$	$\rho = 0.3$	0.75	0.89	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.32	0.47	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.12	0.20	0.88	0.89	1.00	1.00

Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.005$

		$E(N_{cases}) = 500$					
		$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$		$\beta_1^\star = \log(2)$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.03	0.04	0.03	0.04	0.04	0.04
	$\rho = 0.3$	0.04	0.06	0.05	0.05	0.05	0.06
	$\rho = 0.5$	0.03	0.05	0.04	0.05	0.05	0.05
	$\rho = 0.7$	0.04	0.06	0.06	0.06	0.06	0.06
$p = 0.1$	$\rho = 0.1$	0.98	0.99	1.00	1.00	1.00	1.00
	$\rho = 0.3$	0.29	0.37	0.90	0.90	0.99	1.00
	$\rho = 0.5$	0.15	0.20	0.47	0.48	0.71	0.71
	$\rho = 0.7$	0.09	0.12	0.24	0.25	0.34	0.35
$p = 0.2$	$\rho = 0.3$	0.78	0.84	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.32	0.39	0.92	0.92	1.00	1.00
	$\rho = 0.7$	0.14	0.18	0.58	0.59	0.83	0.84
$p = 0.3$	$\rho = 0.3$	0.97	0.99	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.59	0.66	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.24	0.31	0.88	0.89	0.99	0.99

Cox model with $P(\delta = 1) = 0.005$

		$E(N_{cases}) = 500$					
		$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$		$\beta_1^\star = \log(2)$	
		$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test	$\hat{p}$ test	$\hat{d}$ test
$p = 0.0$	$\rho = 0.1$	0.04	0.06	0.04	0.05	0.05	0.05
	$\rho = 0.3$	0.04	0.06	0.05	0.05	0.05	0.06
	$\rho = 0.5$	0.03	0.04	0.04	0.05	0.03	0.03
	$\rho = 0.7$	0.03	0.05	0.04	0.05	0.04	0.04
$p = 0.1$	$\rho = 0.1$	0.95	0.99	1.00	1.00	1.00	1.00
	$\rho = 0.3$	0.30	0.38	0.88	0.88	0.99	0.99
	$\rho = 0.5$	0.12	0.17	0.45	0.46	0.71	0.72
	$\rho = 0.7$	0.07	0.10	0.22	0.22	0.39	0.39
$p = 0.2$	$\rho = 0.3$	0.73	0.82	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.30	0.37	0.92	0.93	1.00	1.00
	$\rho = 0.7$	0.14	0.19	0.53	0.54	0.84	0.84
$p = 0.3$	$\rho = 0.3$	0.95	0.97	1.00	1.00	1.00	1.00
	$\rho = 0.5$	0.60	0.67	1.00	1.00	1.00	1.00
	$\rho = 0.7$	0.27	0.34	0.86	0.87	0.99	0.99

Table 4: Coverage rates (CI-RATE) and lengths (CI-LEN) of trimmed untransformed and logit transformed-based (Trans) confidence intervals for the mediation proportion under the identity and logit link functions and the Cox model

Identity link function ($Y \sim Normal$)

			$n = 1000$				$n = 5000$			
			$\beta_1^\star = 0.3$		$\beta_1^\star = 0.5$		$\beta_1^\star = 0.3$		$\beta_1^\star = 0.5$	
			Trim	Trans	Trim	Trans	Trim	Trans	Trim	Trans
$p = 0.1$	$\rho = 0.1$	CI-RATE	0.96	0.96	0.95	0.94	0.95	0.96	0.95	0.95
		CI-LEN	0.13	0.14	0.12	0.13	0.06	0.06	0.05	0.05
	$\rho = 0.3$	CI-RATE	0.94	0.96	0.94	0.96	0.95	0.95	0.94	0.94
		CI-LEN	0.14	0.16	0.09	0.09	0.06	0.06	0.04	0.04
	$\rho = 0.5$	CI-RATE	0.91	0.88	0.94	0.95	0.95	0.95	0.95	0.95
		CI-LEN	0.21	0.36	0.14	0.17	0.11	0.11	0.07	0.07
	$\rho = 0.7$	CI-RATE	0.73	0.73	0.88	0.88	0.95	0.95	0.95	0.95
		CI-LEN	0.33	0.55	0.21	0.35	0.17	0.23	0.11	0.11
$p = 0.3$	$\rho = 0.3$	CI-RATE	0.95	0.96	0.95	0.95	0.95	0.95	0.94	0.95
		CI-LEN	0.20	0.20	0.14	0.14	0.09	0.09	0.06	0.06
	$\rho = 0.5$	CI-RATE	0.95	0.97	0.95	0.95	0.95	0.95	0.95	0.95
		CI-LEN	0.28	0.27	0.17	0.17	0.12	0.12	0.07	0.07
	$\rho = 0.7$	CI-RATE	0.98	0.98	0.97	0.97	0.94	0.94	0.94	0.94
		CI-LEN	0.42	0.42	0.26	0.25	0.19	0.19	0.11	0.11
$p = 0.5$	$\rho = 0.5$	CI-RATE	0.95	0.97	0.95	0.95	0.95	0.95	0.95	0.95
		CI-LEN	0.33	0.32	0.20	0.20	0.14	0.14	0.09	0.09
	$\rho = 0.7$	CI-RATE	0.98	0.98	0.96	0.96	0.95	0.95	0.95	0.95
		CI-LEN	0.46	0.43	0.28	0.27	0.20	0.20	0.12	0.12

Logit link function ($Y \sim Ber$) with $P(Y = 1) = 0.005$

			$E(N_{cases}) = 500$				$E(N_{cases}) = 1000$			
			$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$		$\beta_1^\star = \log(1.25)$		$\beta_1^\star = \log(1.5)$	
			Trim	Trans	Trim	Trans	Trim	Trans	Trim	Trans
$p = 0.1$	$\rho = 0.1$	CI-RATE	0.95	0.96	0.94	0.98	0.95	0.96	0.95	0.95
		CI-LEN	0.08	0.08	0.15	0.16	0.06	0.06	0.04	0.04
	$\rho = 0.3$	CI-RATE	0.96	0.95	0.90	0.87	0.95	0.95	0.95	0.96
		CI-LEN	0.18	0.24	0.29	0.48	0.14	0.16	0.10	0.10
	$\rho = 0.5$	CI-RATE	0.88	0.82	0.77	0.72	0.93	0.89	0.96	0.93
		CI-LEN	0.28	0.46	0.46	0.69	0.22	0.36	0.17	0.23
	$\rho = 0.7$	CI-RATE	0.71	0.71	0.59	0.59	0.76	0.76	0.85	0.85
		CI-LEN	0.44	0.69	0.72	0.86	0.35	0.58	0.25	0.45
$p = 0.3$	$\rho = 0.3$	CI-RATE	0.94	0.96	0.94	0.98	0.94	0.95	0.95	0.96
		CI-LEN	0.25	0.25	0.45	0.43	0.18	0.18	0.13	0.13
	$\rho = 0.5$	CI-RATE	0.96	0.99	0.95	0.97	0.95	0.97	0.95	0.96
		CI-LEN	0.37	0.36	0.61	0.63	0.28	0.27	0.20	0.20
	$\rho = 0.7$	CI-RATE	0.97	0.97	0.88	0.88	0.98	0.98	0.97	0.97
		CI-LEN	0.55	0.58	0.79	0.83	0.44	0.44	0.31	0.30
$p = 0.5$	$\rho = 0.5$	CI-RATE	0.96	0.97	0.90	0.94	0.96	0.97	0.94	0.96
		CI-LEN	0.44	0.42	0.67	0.63	0.31	0.30	0.23	0.22
	$\rho = 0.7$	CI-RATE	0.99	0.99	0.91	0.91	0.98	0.98	0.97	0.97
		CI-LEN	0.60	0.57	0.83	0.82	0.47	0.45	0.34	0.33

Cox model with $P(\delta = 1) = 0.005$

			$E(N_{cases}) = 500$				$E(N_{cases}) = 1000$			
			$\beta_1^\star = log(1.25)$		$\beta_1^\star = log(1.5)$		$\beta_1^\star = log(1.25)$		$\beta_1^\star = log(1.5)$	
			Trim	Trans	Trim	Trans	Trim	Trans	Trim	Trans
$p = 0.1$	$\rho = 0.1$	CI-RATE	0.94	0.95	0.93	0.98	0.94	0.96	0.93	0.94
		CI-LEN	0.08	0.08	0.17	0.18	0.07	0.07	0.05	0.05
	$\rho = 0.3$	CI-RATE	0.96	0.96	0.90	0.86	0.95	0.96	0.94	0.96
		CI-LEN	0.18	0.24	0.29	0.48	0.14	0.16	0.10	0.11
	$\rho = 0.5$	CI-RATE	0.88	0.82	0.77	0.73	0.93	0.89	0.94	0.92
		CI-LEN	0.28	0.47	0.46	0.71	0.22	0.37	0.17	0.22
	$\rho = 0.7$	CI-RATE	0.68	0.68	0.62	0.62	0.76	0.76	0.84	0.84
		CI-LEN	0.44	0.68	0.73	0.86	0.34	0.59	0.26	0.43
$p = 0.3$	$\rho = 0.3$	CI-RATE	0.96	0.96	0.93	0.98	0.95	0.95	0.97	0.97
		CI-LEN	0.26	0.25	0.47	0.44	0.20	0.19	0.14	0.14
	$\rho = 0.5$	CI-RATE	0.96	0.99	0.94	0.98	0.94	0.97	0.95	0.96
		CI-LEN	0.37	0.36	0.61	0.63	0.29	0.28	0.20	0.20
	$\rho = 0.7$	CI-RATE	0.98	0.98	0.89	0.89	0.97	0.97	0.98	0.98
		CI-LEN	0.54	0.57	0.78	0.82	0.45	0.45	0.31	0.31
$p = 0.5$	$\rho = 0.5$	CI-RATE	0.94	0.95	0.91	0.94	0.96	0.97	0.95	0.96
		CI-LEN	0.44	0.42	0.69	0.64	0.34	0.33	0.24	0.23
	$\rho = 0.7$	CI-RATE	0.99	0.99	0.90	0.90	0.98	0.98	0.97	0.97
		CI-LEN	0.61	0.57	0.83	0.81	0.48	0.46	0.34	0.33

Table 5: 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). When $\hat{p} \notin [0,1]$, inference for the mediation proportion was not carried out. $\hat{RR}_{total} = \exp(\hat{\beta}^{\star})$, $\hat{RR}_{direct} = \exp(\hat{\beta})$.

Risk factor	$\hat{\beta}^{\star}$ ($\hat{RR}_{total}$)	p-value	$\hat{\beta}$ ($\hat{RR}_{direct}$)	$\hat{p}$	CrudeCI95%	TrimCI95%	TransCI95%	p-value, $\hat{p}$ test	p-value, $\hat{d}$ test
Personal history of benign breast disease	0.35 (1.42)	< 0.001	0.25 (1.28)	0.30	0.10–0.51	0.10–0.51	0.14–0.53	0.002	$< 10^{-6}$
Family history of breast cancer	0.42 (1.52)	0.01	0.42 (1.52)	0.004	-0.10–0.11	0.00–0.11	0.00–1.00	0.47	0.47
Adolescent somatotype† Per 3 unit increase	-0.34 (0.72)	0.02	-0.12 (0.88)	0.63	0.05–1.20	0.05–1.00	0.13–0.95	0.02	$< 10^{-7}$
BMI at age 18† Per 5 unit increase	-0.23 (0.79)	0.02	-0.05 (0.95)	0.78	0.06–1.50	0.06–1.00	0.05–1.00	0.02	$< 10^{-8}$
Age at first birth‡ Per 5 year increase	0.15 (1.17)	0.03	0.15 (1.16)	0.03	-0.09–0.15	0.00–0.15	0.00–0.66	0.31	0.30
Age at menarche Per 2 year increase	-0.16 (0.86)	0.03	-0.18 (0.84)	N/A	N/A	N/A	N/A	N/A	N/A
Height Per 3 inch increase	0.13 (1.14)	0.03	0.14 (1.14)	N/A	N/A	N/A	N/A	N/A	N/A

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

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

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