Statistica Sinica **34** (2024), 1045-1066
doi:https://doi.org/10.5705/ss.202021.0373

INTEGRATING INCOMPLETE DATA FOR MEDIATION ANALYSIS

Andriy Derkach*[1], Joshua N. Sampson[2] and Ruth M. Pfeiffer[2]

[1]*Memorial Sloan-Kettering Cancer Center and* [2]*National Cancer Institute*

Abstract: Mediation analysis examines the relationships between an exposure, a mediator, and an outcome. Although many approaches are available for performing such analyses they all require access to a single complete data set that contains the three key variables: outcome, exposure, and mediator. Here, we propose semiparametric methods for mediation analysis to estimate the standard causal parameters (direct and indirect effects) by combining information from several incomplete data sets, each containing only two of the three key variables. Importantly, our methods also handle scenarios in which only summary statistics based on those data sets are available. The resulting estimates of the causal parameters are asymptotically unbiased and normally distributed. We evaluate the performance of our methods in finite samples using simulations, and quantify the loss in efficiency from the lack of a complete data set with all three variables. We then apply proposed method to determine whether the number of terminal duct lobular units in the breast mediate the relationship between a polygenic risk score and breast cancer risk.

Key words and phrases: Data integration, direct and indirect effects, semiparametric likelihood, summary level information.

1. Introduction

Mediation analysis has become a popular statistical tool for understanding the relationship between an exposure (E), a mediator (M), and an outcome (Y). Many advanced methods for mediation analysis have recently been developed that handle almost any situation (e.g., Derkach et al. (2019); Daniels et al. (2012); Huang (2019); Cheng et al. (2018); Zeng et al. (2021); Huang and Cai (2016)). However, all available methods have the critical limitation that they require the three relevant variables to be measured in a single common data set. Here, we develop methods for mediation analysis when we have three "incomplete" data sets, each containing only two of the three variables (i.e., E and M, M and Y, E and Y).

In our motivating study (Bodelon et al. (2020)), we wanted to understand how genetic risk factors, summarized by a polygenic risk score (PRS, E), influenced a woman's breast cancer risk (Y). The effect of the PRS may be

*Corresponding author.

mediated by the number of terminal duct lobular units (TDLUs, M). TDLUs are epithelial structures that produce milk during lactation. We measured the PRS and the number of TDLUs in a cohort of women, yielding a data set with two of the three needed variables (E and M, but not Y). However, published summary statistics (odds ratios, ORs) for the association of PRS and breast cancer risk (E and Y) (Mavaddat et al. (2019)) and between TDLUs and breast cancer risk (M and Y) (Figueroa et al. (2014)) are available. Thus, we aim to perform a mediation analysis by combining information from these three incomplete data sets.

Here, we develop methods for performing a mediation analysis in three scenarios. For all scenarios, we assume we have a data set containing individual-level data on E and M. In the first scenario, we have two summary statistics (e.g., published ORs) that capture the relationships between E and Y and between M and Y, respectively. In the second scenario, we have a summary statistic describing one relationship (either between E and Y, or between M and Y), and a data set with individual-level measurements for evaluating the second relationship. In the third scenario, we have two additional data sets, each containing individual-level data on two of the three variables. For all scenarios, we assume that information on a common set of covariates is available for all three data sources.

We highlight some key features of our proposed methods. The methods can be applied to outcomes that follow any distribution in the exponential family, do not require parametric assumptions about the joint distribution of E and M, and accommodate differences in covariate adjustments. They can handle interactions between M and a categorical E when the effect of M on Y is measured in subgroups defined by categories of E. They can also be extended to include multiple exposures and mediators.

Our proposed methods extend approaches developed to handle studies with missing data. We draw heavily on ideas from two-phase (e.g., case-cohort) designs, where some variables are measured on an entire cohort, and then a limited number of "expensive" variables (e.g., biomarkers) are measured only on a small sub-sample of individuals. Specifically, we build on methods that use a semiparametric maximum likelihood (e.g., Lin and Zeng (2006); Breslow and Holubkov (1997)). We also draw from methods that calibrate models, using published summary-level statistics from large studies, using a constrained maximum likelihood estimation (Chatterjee et al. (2016); Zhang et al. (2020); Kundu, Tang and Chatterjee (2019)). However, these methods all assume that at least one data set contains measurements on all primary variables.

The remainder of the paper is organized as follows. We first describe the statistical methods and discuss the theoretical properties of the resulting estimates (Section 2). We then study the estimates in finite samples using simulations (Section 3), and analyze breast cancer data (Section 4), Section 5 concludes the paper.

2. Methods

2.1. Overview

We first describe the model and target parameters. We assume that given an exposure E, a mediator M, and covariates $\boldsymbol{X} = (X_1, \ldots, X_k)$, the distribution of the outcome Y belongs to an exponential family,

$$f(Y|M, E, \boldsymbol{X}; \theta, \psi) = \exp\left[\frac{Y\theta - b(\theta)}{a(\psi)} + c(Y, \psi)\right], \tag{2.1}$$

with

$$\theta = \alpha + \beta M + \gamma E + \boldsymbol{\delta}'\boldsymbol{X}. \tag{2.2}$$

We let $\boldsymbol{\eta} = (\alpha, \beta, \gamma, \boldsymbol{\delta})'$ denote all parameters in (2.2), including a vector of covariate effects, $\boldsymbol{\delta} \in \mathbb{R}^k$.

We partition the total effect (TE) of changing the exposure value from $E = e$ to $E = e'$ into a natural direct effect (NDE) and a natural indirect effect (NIE), while controlling for $\boldsymbol{X}$, as follows:

$$TE_x = \mathbb{E}\left[Y\{e', M(e')\} - Y\{e, M(e)\} \middle| \boldsymbol{X} = \boldsymbol{x}\right] = NDE_x + NIE_x,$$

where

$$NDE_x = \mathbb{E}\left[Y\{e', M(e)\} - Y\{e, M(e)\} \middle| \boldsymbol{X} = \boldsymbol{x}\right] \text{ and} \tag{2.3}$$

$$NIE_x = \mathbb{E}\left[Y\{e', M(e')\} - Y\{e', M(e)\} \middle| \boldsymbol{X} = \boldsymbol{x}\right]. \tag{2.4}$$

Under models (2.1) and (2.2) and the assumptions outlined in Imai, Keele and Tingley (2010); Imai, Keele and Yamamoto (2011), the expectations are calculated using the conditional distribution

$$\mathbb{E}\left[Y\{e, M(e')\} | \boldsymbol{X} = \boldsymbol{x}\right] = \int_M b'(\alpha + \beta m + \gamma e + \boldsymbol{\delta}'\boldsymbol{x}) dF_{M|E=e', \boldsymbol{X}=\boldsymbol{x}}(m).$$

There are many parametric, semiparametric and nonparametric approaches for estimating the TE, NDE, and NIE when individual-level data on Y, M, E, and $\boldsymbol{X}$ are available on all subjects in a study (e.g., Daniels et al. (2012); Derkach et al. (2019). Here, we describe approaches for estimating these parameters when we have multiple incomplete data sets. Specifically, we consider three scenarios that are defined by the available information. For all scenarios, we assume that we have a data set with individual level information on $(M, E, \boldsymbol{X})$, and that all three sources of information are based on samples from the same underlying population. The three scenarios are as follows:

1. We have estimates of the associations between M and Y and between E and Y, with both estimates adjusted for the common set of covariates $\boldsymbol{X}$.

2. We have an estimate of the association between M and Y, adjusted for $\boldsymbol{X}$, and a data set with individual-level data on $(M, E, \boldsymbol{X})$, or, alternatively, we have an estimate of the association between E and Y, adjusted for $\boldsymbol{X}$, and individual-level data on $(Y, M, \boldsymbol{X})$.

3. We have two data sets, one with individual-level measurements on $(Y, E, \boldsymbol{X})$, and one with individual-level data on $(Y, M, \boldsymbol{X})$.

We comment on scenarios in which the data sets have different sets of covariates in Table 1 in Section 2.9. For brevity, we denote the full predictor set by $\mathscr{D} = (M, E, \mathbf{X})$, and a specific realization by $\mathscr{d} = (m, e, \boldsymbol{x})$.

2.2. Estimation based on two marginal estimates: Scenario 1

In this scenario, we have a data set with individual-level measurements $\mathscr{D}_i = (M_i, E_i, \mathbf{X}_i)$, for $i = 1, \ldots, N_1$. We also have an estimate of the marginal association between M and Y, adjusted for $\mathbf{X}$. In other words, a prior study collected data on $(M, Y, \mathbf{X})$, assuming

$$f(Y|M, \mathbf{X}; \theta_M, \psi_M) = \exp\left[\frac{Y\theta_M - b(\theta_M)}{a(\psi_M)} + c(Y, \psi_M)\right], \tag{2.5}$$

with $\theta_M = \alpha_M + \beta_M M + \boldsymbol{\delta}_M'\mathbf{X}$ and reported $\hat{\boldsymbol{\eta}}_M = (\hat{\alpha}_M, \hat{\beta}_M, \hat{\boldsymbol{\delta}}_M)$. We also have an estimate of the marginal association between E and Y, adjusted for $\boldsymbol{X}$. In other words, another study collected data on $(E, Y, \mathbf{X})$, assuming

$$f(Y|E, \boldsymbol{X}; \theta_E, \psi_E) = \exp\left[\frac{Y\theta_E - b(\theta_E)}{a(\psi_E)} + c(Y, \psi_E)\right], \tag{2.6}$$

with $\theta_E = \alpha_E + \gamma_E E + \boldsymbol{\delta}_E'\boldsymbol{X}$ and reported $\hat{\boldsymbol{\eta}}_E = (\hat{\alpha}_E, \hat{\gamma}_E, \hat{\boldsymbol{\delta}}_E)$. For now, we assume that the data sets are large so that both estimates $\hat{\boldsymbol{\eta}}_M$ and $\hat{\boldsymbol{\eta}}_E$ are close to the true values $\boldsymbol{\eta}_M = (\alpha_M, \beta_M, \boldsymbol{\delta}_M)$ and $\boldsymbol{\eta}_E = (\alpha_E, \gamma_E, \boldsymbol{\delta}_E)$, respecyivly. We now propose a new semiparametric method to estimate parameters $\boldsymbol{\eta}$ for the joint model given in (2.1) and (2.2), using $\boldsymbol{\eta}_M$, $\boldsymbol{\eta}_E$,and $\mathscr{D}_i$, for $i = 1, \ldots, N_1$.

Letting ∇ denote the gradient operator that yields the vector of partial derivatives, the score vectors for the working models (i.e., the models that do not contain E, M, and Y) of the form (2.5) and (2.6) are

$$\begin{aligned}\boldsymbol{U}_M(\boldsymbol{\eta}_M) &= \nabla_{\boldsymbol{\eta}_M} \log\{f(Y|M, \boldsymbol{X}; \theta_M, \psi_M)\},\\ \boldsymbol{U}_E(\boldsymbol{\eta}_E) &= \nabla_{\boldsymbol{\eta}_E} \log\{f(Y|E, \boldsymbol{X}; \theta_E, \psi_E)\}.\end{aligned}$$

Following Chatterjee et al. (2016), under mild conditions White (1982), the expectations of the score vectors under the true model (2.1) and (2.2) can be used to convert the external marginal estimates $\boldsymbol{\eta}_M$ and $\boldsymbol{\eta}_E$ into a system of

equations with unique solutions corresponding to the true parameters $\boldsymbol{\eta}$,

$$\mathbb{E}\left\{\boldsymbol{U}_M(\boldsymbol{\eta}_M);\boldsymbol{\eta}\right\} = \int_{\mathscr{D}} \left\{ \int_Y \boldsymbol{U}_M(\boldsymbol{\eta}_M) f(Y|\mathscr{D};\theta,\psi) dY \right\} dF(\mathscr{D}) = 0 \tag{2.7}$$

$$\mathbb{E}\left\{\boldsymbol{U}_E(\boldsymbol{\eta}_E);\boldsymbol{\eta}\right\} = \int_{\mathscr{D}} \left\{ \int_Y \boldsymbol{U}_E(\boldsymbol{\eta}_E) f(Y|\mathscr{D};\theta,\psi) dY \right\} dF(\mathscr{D}) = 0, \tag{2.8}$$

where $f(Y|\mathscr{D};\theta,\psi)$ corresponds to the full model (2.1). The system of equations (2.7 and 2.8) does not require that the working models (2.5) and (2.6) and the full model (2.1) use the same link functions, but does assume that the joint distribution of $(Y, M, E, \boldsymbol{X})$ is the same in the two studies.

For the canonical link, that is, $\theta = \theta(\boldsymbol{\eta})$, the system of equations (2.7) is

$$\begin{aligned}
\int_{\mathscr{D}} \left\{ b'(\alpha + \beta M + \gamma E + \boldsymbol{\delta}'\boldsymbol{X}) - b'(\alpha_M + \beta_M M + \boldsymbol{\delta}'_M \boldsymbol{X}) \right\} dF(\mathscr{D}) = 0, \\
\int_{\mathscr{D}} \left\{ b'(\alpha + \beta M + \gamma E + \boldsymbol{\delta}'\boldsymbol{X}) - b'(\alpha_M + \beta_M M + \boldsymbol{\delta}'_M \boldsymbol{X}) \right\} M dF(\mathscr{D}) = 0, \\
\int_{\mathscr{D}} \left\{ b'(\alpha + \beta M + \gamma E + \boldsymbol{\delta}'\boldsymbol{X}) - b'(\alpha_M + \beta_M M + \boldsymbol{\delta}'_M \boldsymbol{X}) \right\} \boldsymbol{X} dF(\mathscr{D}) = 0.
\end{aligned}$$

We obtain three equations based on (2.8) similarly. The intercept α and the covariate-specific parameters $\boldsymbol{\delta}$ are present in equations (2.7) based on $\boldsymbol{U}_M$, and in (2.8) based on $\boldsymbol{U}_E$. To eliminate this over-determination we add the two score equations for α and $\boldsymbol{\delta}$, and estimate $\boldsymbol{\eta} = (\alpha, \beta, \gamma, \boldsymbol{\delta})$ by solving the system of equations

$$\begin{aligned}
S(\alpha) &= \int_{\mathscr{D}} \left\{ b'(\theta) - \frac{b'(\theta_M)}{2} - \frac{b'(\theta_E)}{2} \right\} dF(\mathscr{D}) = 0 \\
S(\beta) &= \int_{\mathscr{D}} \left\{ b'(\theta) - b'(\theta_M) \right\} M dF(\mathscr{D}) = 0 \\
S(\gamma) &= \int_{\mathscr{D}} \left\{ b'(\theta) - b'(\theta_E) \right\} E dF(\mathscr{D}) = 0. \\
S(\delta) &= \int_{\mathscr{D}} \left\{ b'(\theta) - \frac{b'(\theta_M)}{2} - \frac{b'(\theta_E)}{2} \right\} \boldsymbol{X} dF(\mathscr{D}) = 0,
\end{aligned} \tag{2.9}$$

with $\theta = \alpha + \beta M + \gamma E + \boldsymbol{\delta}'\boldsymbol{X}$, $\theta_M = \alpha_M + \beta_M M + \boldsymbol{\delta}'_M \boldsymbol{X}$, and $\theta_E = \alpha_E + \gamma_E E + \boldsymbol{\delta}'_E \boldsymbol{X}$. We now illustrate the approach for a special case.

Example 1. Let Y given M, E, and $\boldsymbol{X}$ be normally distributed. Then, the system of equations (2.9), written in matrix form, simplifies to

$$\begin{pmatrix} 1 & \mathbb{E}(M) & \mathbb{E}(E) & \mathbb{E}(\mathbf{X}') \\ \mathbb{E}(M) & \mathbb{E}(M)^2 & \mathbb{E}(ME) & \mathbb{E}(M\boldsymbol{X}') \\ \mathbb{E}(E) & \mathbb{E}(ME) & \mathbb{E}(E^2) & \mathbb{E}(E\boldsymbol{X}') \\ \mathbb{E}(\mathbf{X}) & \mathbb{E}(M\mathbf{X}) & \mathbb{E}(E\mathbf{X}) & \mathbb{E}(\mathbf{X}\mathbf{X}') \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \boldsymbol{\delta} \end{pmatrix} = \begin{pmatrix} [\mathbb{E}(\theta_E) + \mathbb{E}(\theta_M)]/2 \\ \mathbb{E}(\theta_M M) \\ \mathbb{E}(\theta_E E) \\ [\mathbb{E}(\theta_E \mathbf{X}) + \mathbb{E}(\theta_M \mathbf{X})]/2 \end{pmatrix}.$$

The left-most matrix is the Fisher information matrix $\mathbb{I}$ under a normal model with mean (2.2), and there is a unique solution for $\boldsymbol{\eta} = (\alpha, \beta, \gamma, \boldsymbol{\delta})$ if and only if $\mathbb{I}$ has full rank (i.e., is invertible).

The normal example points to the proof of the proposition for outcomes Y with distributions in the exponential family which we state next.

Proposition 1. *Under the model defined by* (2.1) *and* (2.2)*, the system of equations* (2.9) *has a unique solution that is equal to the true parameters* $\boldsymbol{\eta} = (\alpha, \beta, \gamma, \boldsymbol{\delta})'$ *if and only if the Fisher information matrix* $\mathbb{I}$ *has full rank.*

Proof. To demonstrate that only the true parameters $\boldsymbol{\eta}$ are solutions to the system of equations (2.9), we apply the inverse function theorem "Theorems about Differentiable Functions" (Allendoerfer (1974)). This theorem states that the corresponding objective function to the system of equations (2.9) is strictly convex with a unique minimum iff the Jacobian with respect to $\boldsymbol{\eta}$ is positive definite.

Let $\mathbf{S} = (S(\alpha), S(\beta), S(\gamma), S(\boldsymbol{\delta}))'$, with S given in (2.9). The Jacobian matrix with respect to $\boldsymbol{\eta}$ is

$$J = (\nabla_{\boldsymbol{\eta}} \mathbf{S}) = \begin{pmatrix} \mathbb{E}(b''(\theta)) & \mathbb{E}(b''(\theta)\mathscr{D}') \\ \mathbb{E}(b''(\theta)\mathscr{D}) & \mathbb{E}(b''(\theta)\mathscr{D}\mathscr{D}') \end{pmatrix},$$

which is the Fisher information matrix, $\mathbb{I}$. Thus, the solution $\boldsymbol{\eta}$ is unique if and only if $\mathbb{I}$ has full rank.

Remark 1. We assume that the intercepts α_M and α_E are provided, and that the true intercept α is obtained from the system of equations (2.9). For some special cases of distributions of the form of (2.1) and (2.2), one can estimate β, γ, and $\boldsymbol{\delta}$, while ignoring the intercept. Let $\mathbb{E}(M) = \mathbb{E}(E) = \mathbb{E}(X_1) = \cdots = \mathbb{E}(X_k) = 0$. Then, for example, when Y given E, M and $\boldsymbol{X}$ is normally distributed, $\alpha = \alpha_E = \alpha_M$ and thus all intercepts can be set to zero. When Y given E, M, and $\mathbf{X}$ has a logistic distribution in which β, γ, and δ_j, for $j = 1, \ldots, k$, are small, then $\alpha \approx \alpha_E \approx \alpha_M$ and $\alpha \approx \log\{P(Y = 1)/P(Y = 0)\}$, where $P(Y = 1)$ is the prevalence of the outcome Y in the source population.

Remark 2. Estimates $\boldsymbol{\eta}_M$ and $\boldsymbol{\eta}_E$ from retrospective case-control studies can be used to obtain consistent estimates of $(\beta, \gamma, \boldsymbol{\delta})$, even though the data in such studies do not follow the population distribution of $\mathscr{D}_i$ (Carroll, Wang and Wang (1995)). However, α is not estimable. Here, we propose setting $\alpha = \log\{P(Y = 1)/P(Y = 0)\}$ when β, γ, and $P(Y = 1)$ are small. In our simulations, we show the robustness of this approach for rare outcomes.

To obtain a solution for the system of equations (2.9), one needs to specify a joint distribution $F(\mathscr{D})$. We estimate F using the empirical distribution based on the individual-level observations $\mathscr{D}_i$, characterized by $\hat{F}(\mathscr{d}) =$

$(1/N_1)\sum_{j=1}^{N_1} I\{\mathscr{D}_j \le \mathscr{d}\}$, where I is the indicator function and $dF(\mathscr{d}_i)$ is the point mass for the ith observation. We thus obtain

$$\begin{aligned}
S_\alpha(\hat{F}) &= \sum_{i=1}^{N_1}\left\{b'(\theta_i) - \frac{b'(\theta_{Mi})}{2} - \frac{b'(\theta_{Ei})}{2}\right\} = 0,\\
S_\beta(\hat{F}) &= \frac{1}{N_1}\sum_{i=1}^{N_1} m_i\left\{b'(\theta_i) - b'(\theta_{Mi})\right\} = 0,\\
S_\gamma(\hat{F}) &= \frac{1}{N_1}\sum_{i=1}^{N_1} e_i\left\{b'(\theta_i) - b'(\theta_{Ei})\right\} = 0,\\
S_\delta(\hat{F}) &= \frac{1}{N_1}\sum_{i=1}^{N_1} \boldsymbol{x}_i\left\{b'(\theta_i) - \frac{b'(\theta_{Mi})}{2} - \frac{b'(\theta_{Ei})}{2}\right\} = \mathbf{0}.
\end{aligned} \tag{2.10}$$

The equations in (2.10) are of the form $\sum_{i=1}^{N_1}\mathbf{S}(\mathscr{D}_i,\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^*) = 0$, where $\hat{\boldsymbol{\eta}}^* = (\hat{\boldsymbol{\eta}}_E,\hat{\boldsymbol{\eta}}_M)$ are estimates of $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_E,\boldsymbol{\eta}_M)$ from studies with sample sizes N_2 and N_3.

Proposition 2. *Assume that $N_i \to \infty$, for $i = 1,2,3$, such that $N_k/N_1 \to \rho_k$ (with $0 < \rho_k < \infty$) for $k = 2,3$, and as $N_1 \to \infty$, $\hat{\boldsymbol{\eta}}^*$ converges to a normal distribution, $\sqrt{N_1}(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*) \underset{d}{\to} N(\mathbf{0},\Sigma_{\eta^*})$. Then, under the model in* (2.1) *and* (2.2)*, the solution $\hat{\boldsymbol{\eta}} = (\hat{\alpha},\hat{\beta},\hat{\gamma},\boldsymbol{\delta})'$ of* (2.10) *satisfies*

$$\lim_{N_1\to\infty}\sqrt{N_1}\,(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \underset{d}{\to} N\left(\mathbf{0}, J^{-1}B(\boldsymbol{\eta})J^{-1} + J^{-1}\Omega\Sigma_{\eta^*}\Omega' J^{-1}\right), \tag{2.11}$$

where $B(\boldsymbol{\eta}) = \mathbb{E}\{\mathbf{S}(\mathscr{D},\boldsymbol{\eta},\boldsymbol{\eta}^*)\mathbf{S}(\mathscr{D},\boldsymbol{\eta},\boldsymbol{\eta}^*)'\}$, $\Omega = \mathbb{E}\{\nabla_{\boldsymbol{\eta}^*}\mathbf{S}(\mathscr{D},\boldsymbol{\eta},\boldsymbol{\eta}^*)\}$, and $J = \mathbb{E}\{\nabla_{\boldsymbol{\eta}}\mathbf{S}(\mathscr{D},\boldsymbol{\eta},\boldsymbol{\eta}^*)\}$.

Proof. By the Taylor expansion around $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$,

$$0 = \sum_{i=1}^{N_1}\mathbf{S}(\mathscr{D}_i,\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^*) = \sum_{i=1}^{N_1}\mathbf{S}(\mathscr{D}_i,\hat{\boldsymbol{\eta}},\boldsymbol{\eta}^*) + \sum_{i=1}^{N_1}\nabla_{\boldsymbol{\eta}^*}\mathbf{S}(\mathscr{D}_i,\hat{\boldsymbol{\eta}},\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*) + o_P(\sqrt{N_1})$$

and

$$\begin{aligned}
o_P(\sqrt{N_1}) = &\sum_{i=1}^{N_1}\mathbf{S}(\mathscr{D}_i,\boldsymbol{\eta},\boldsymbol{\eta}^*) + \sum_{i=1}^{N_1}\nabla_{\boldsymbol{\eta}}\mathbf{S}(\mathscr{D}_i,\boldsymbol{\eta},\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\\
&+ \sum_{i=1}^{N_1}\nabla_{\boldsymbol{\eta}^*}\mathbf{S}(\mathscr{D}_i,\boldsymbol{\eta},\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*).
\end{aligned}$$

In addition, we observe that, asymptotically,

$$\begin{bmatrix}\frac{1}{\sqrt{N_1}}\sum_i \mathbf{S}(\mathscr{D}_i,\boldsymbol{\eta},\boldsymbol{\eta}^*)\\ \sqrt{N_1}(\hat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^*)\end{bmatrix} = N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}B(\boldsymbol{\eta}) & 0\\ 0 & \Sigma_{\eta^*}\end{bmatrix}\right). \tag{2.12}$$

Applying Theorem 1 of Yuan and Jennrich (2000) yields the result in (2.11).

2.3. Estimation based on one incomplete data set and a marginal estimate: Scenario 2

We now assume that in addition to the study with individual-level data on $\mathscr{D}_i$, for $i = 1, \ldots, N_1$, we have another (incomplete) data set with individual-level data on $(Y_j, E_j, \boldsymbol{X}_j)$, for $j = 1, \ldots, N_2$, and marginal estimates $\hat{\boldsymbol{\eta}}_M$. The scenario with individual-level data on $(Y, M, \boldsymbol{X})$ and marginal estimates $\hat{\boldsymbol{\eta}}_E$ can be handled following the same approach.

Similarly to Chatterjee et al. (2016), we construct the observed-data likelihood

$$L_2(\theta, \psi) = \prod_{i=1}^{N_1} f(\mathscr{D}_i) \prod_{k=1}^{N_2} \int_M f(Y_k | \mathscr{D}_k; \theta, \psi) f(\mathscr{D}_k) dM_k, \tag{2.13}$$

under the constraint $\mathbb{E}\{U_M(\boldsymbol{\eta}_M); \boldsymbol{\eta}\} = 0$, with $\theta = \alpha + \beta M + \gamma E + \delta' \boldsymbol{X}$. In contrast to scenario 1 with summary statistics only or the problem studied in Chatterjee et al. (2016), $f(\mathscr{D})$ cannot be factored out or estimated only from the first data set without loss of efficiency.

For simplicity, we assume that E and $\boldsymbol{X}$ are discrete, and that all unique values of E and $\boldsymbol{X}$ are observed in both individual-level data sets. Specifically, the distribution of $\mathscr{D}_i$ is given by the point masses $\mathbf{p} = (p_1, \ldots, p_R)$ at the R unique values of $\mathscr{D}$ and $n_r = \sum_{i=1}^{N_1} I\{\mathscr{D}_i = \mathscr{A}_r)\}$, for $r = 1, \ldots, R$. Then, the observed-data likelihood can be written as

$$L_2^s(\theta, \psi) = \prod_{r=1}^{R} p_r^{n_r} \prod_{k=1}^{N_2} \sum_r f(Y_k | \mathscr{A}_r; \theta, \psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r,$$

under the constraint $\sum_{r=1}^{R} \{E(Y|\mathscr{A}_r; \boldsymbol{\eta}) - E(Y|m_r, \boldsymbol{x}_r; \boldsymbol{\eta}_M)\} \mathbf{V}$, where $\mathbf{V}' = (1, m_r, \boldsymbol{x}_r)$. Letting $\boldsymbol{\lambda}$ denote the vector of Lagrange multipliers, the constrained log-likelihood is given by

$$\begin{aligned} LL_2^s(\theta, \psi) = &\sum_{r=1}^{R} n_r \log(p_r) + \sum_{k=1}^{N_2} \log \left\{ \sum_{r=1}^{R} f(Y_k | \mathscr{A}_r; \theta, \psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r \right\} \\ &+ \boldsymbol{\lambda}' N_2 \sum_{r=1}^{R} \{E(Y|\mathscr{A}_r; \boldsymbol{\eta}) - E(Y|m_r, \boldsymbol{x}_r; \boldsymbol{\eta}_M)\} \mathbf{V}. \end{aligned} \tag{2.14}$$

In the Supplementary Material, we propose a computationally efficient and numerically robust expectation-maximization (EM) algorithm that maximizes expression (2.14) as a function of $\boldsymbol{\eta}$ and $\boldsymbol{p}$. Next, we demonstrate uniqueness of the maximum, and establish the consistency and asymptotic normality of the corresponding MLE estimates.

Proposition 3. *Under the model defined by equations* (2.1) *and* (2.2)*, the constrained log-likelihood in* (2.14) *has a unique maximum that is a stationary point.*

Proposition 4. *Let* $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{p}})$ *denote the values that maximize the constrained log-likelihood* (2.14)*. With* $N_1 \to \infty$ *and* $N_k/N_1 \to \rho_k$ *(with* $0 < \rho_k < \infty$*), for* $k = 2, 3$*, under standard regularity conditions (Chatterjee et al. (*2016*)),* $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{p}})$ *is a consistent estimate for* $(\boldsymbol{\eta}, F)$*, and*

$$\lim_{N_2 \to \infty} \sqrt{N_2} \left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \underset{d}{\to} N\left(0, J_\eta\right), \tag{2.15}$$

where J_η *is the asymptotic covariance matrix of* $\boldsymbol{\eta}$ *defined in the Supplementary Material,* S.3.

The proofs for these propositions are given in the Supplementary Material, S.2 and S.3.

2.4. Estimation based on three incomplete data sets: Scenario 3

Here, we assume that we have three incomplete data sets with individual-level data, a study with measures $\mathscr{D}_i$, for $i = 1, \ldots, N_1$, a study with measures $(Y_k, E_k, \boldsymbol{X}_k)$, for $k = 1, \ldots, N_2$, and a study with measures $(Y_j, M_j, \boldsymbol{X}_j)$, for $j = 1, \ldots, N_3$. The observed-data likelihood is

$$\begin{aligned} L_3(\theta, \psi) = &\prod_{i=1}^{N_1} f(\mathscr{D}_i) \prod_{k=1}^{N_2} \int_M f(Y_k | \mathscr{D}_k; \theta, \psi) f(\mathscr{D}_k) dM_k \\ &\prod_{j=1}^{N_3} \int_E f(Y_j | \mathscr{D}_j; \theta, \psi) f(\mathscr{D}_j) dE_j. \end{aligned}$$

Again, $f(\mathscr{D})$ cannot be factored out or estimated from the first data set only without causing a loss of efficiency. We assume that E, M, and $\boldsymbol{X}$ are discrete and all unique values of E, M, and $\boldsymbol{X}$ are present in the data set with measurements on $\mathscr{D}$, and we estimate F nonparametrically with mass points at the unique observed data points. Using the same notation as in the previous sections, the observed-data log-likelihood corresponding to $L_3(\theta, \psi)$ is

$$\begin{aligned} LL_3^s(\theta, \psi) = &\sum_{r=1}^{R} n_r \log(p_r) + \sum_{k=1}^{N_2} \log \left\{ \sum_{r=1}^{R} f(Y_k | \mathscr{A}_r; \theta, \psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r \right\} \\ &+ \sum_{j=1}^{N_3} \log \left\{ \sum_{r=1}^{R} f(Y_j | \mathscr{A}_r; \theta, \psi) I\left[M_j = m_r, \boldsymbol{X}_j = \boldsymbol{x}_r\right] p_r \right\}. \end{aligned} \tag{2.16}$$

In the Supplementary Material, we propose an EM algorithm for maximizing expression (2.16) with respect to $\boldsymbol{\eta}$ and $\boldsymbol{p}$. The next two propositions state the existence of a unique solution and the asymptotic properties of the estimates,

respectively.

Proposition 5. *Under the model defined by equations* (2.1) *and* (2.2)*, the observed-data log-likelihood* (2.16) *has a unique maximum.*

Proposition 6. *Let* $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{p}})$ *maximize the log-likelihood* (2.16)*. Under standard regularity conditions (see Lin and Zeng (2006)), as* $N_1 \to \infty$*,* $N_k/N_1 \to \rho_k$ *(with* $0 < \rho_k < \infty$*), for* $k = 2, 3$*,* $\hat{F} \to F$*, and*

$$\lim_{N_2 \to \infty} \sqrt{N_2} \left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \right) \underset{d}{\to} N \left(0, \mathbb{I}_{\eta}^{-1} \right), \tag{2.17}$$

where $\mathbb{I}_{\eta}^{-1}$ *denotes the inverse information matrix defined in Supplementary Material,* S.3*.*

Our semiparametric method is related to methods for two-phase studies (Breslow and Holubkov (1997)) and response-dependent sampling (e.g. Lin and Zeng (2006)). Proofs of the propositions are provided in the Supplementary Material, S.2 and S.4.

2.5. Estimation under case-control sampling

So far, we have assumed that the distributions of $(Y, E, M, \boldsymbol{X})$ in the three data sources are the same. We now consider the setting in which the two studies with data on a binary Y are conducted using case-control (i.e., outcome-dependent) sampling, and this assumption does not hold. The extension to the scenario in which one study is based on case-control sampling is straight forward. Because the intercept α in a logistic model is not identifiable under case-control sampling, we assume all intercepts α_M, α_E, and α are known. In practice, we propose using $\alpha \approx \alpha_E \approx \alpha_M$ and $\alpha \approx \log\{P(Y = 1)/P(Y = 0)\}$, where $P(Y = 1)$ is the prevalence of Y in the source population. Under scenario 1, the parameters can then be estimated using the approach in Section 2.2. Here, we modify the methods for the other two scenarios.

We start with scenario 2. Let N_2^1 and N_2^0 denote the number of cases and controls, respectively, sampled in the second study, and $P_Y = P(Y = 1) = \int_{\mathscr{D}} P(Y = 1|\mathscr{D}; \boldsymbol{\eta}) dF(\mathscr{D})$ be the marginal probability of Y in the source population. The observed-data likelihood is

$$L_2^R(\theta, \psi) \;=\; \prod_{i=1}^{N_1} f(\mathscr{D}_i) \left\{ \prod_{k=1}^{N_2} \int_M f(Y_k|\mathscr{D}_k; \theta, \psi) f(\mathscr{D}_k) dM_k \right\} P_Y^{-N_2^1} (1 - P_Y)^{-N_2^0},$$

under the constraint $\mathbb{E}\{U_M(\boldsymbol{\eta}_M); \boldsymbol{\eta}\} = 0$. This constraint is based on the score equation for β_M only, as α_M is assumed to be known. Similar to Section 2.3, the constrained semiparametric log-likelihood is

$$LL_2^R(\theta,\psi) = \sum_{r=1}^{R} n_r \log(p_r) + \sum_{k=1}^{N_2} \log\left\{\sum_r f(Y_k|\mathscr{A}_r;\theta,\psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r\right\}$$
$$-N_2^1 \log(P_Y) - N_2^0 \log(1-P_Y)$$
$$+\boldsymbol{\lambda}' N_2 \sum_r \left\{E(Y|\mathscr{A}_r;\boldsymbol{\eta}) - E(Y|m_r,\boldsymbol{x}_r;\boldsymbol{\eta}_M)\right\} \begin{pmatrix} m_r \\ \boldsymbol{x}_r \end{pmatrix} p_r, \tag{2.18}$$

where $P_Y = \sum_r P(Y=1|\mathscr{A}_r,\boldsymbol{x}_r;\theta)p_r$ and $\boldsymbol{\lambda}$ is the vector of Lagrange multiplier. Supplementary Material, S.4 gives an EM algorithm for maximizing the above expression with respect to $(\beta,\gamma,\boldsymbol{\delta})$ and the vector of point masses $\mathbf{p}$. Propositions 3 and 4 still hold under retrospective sampling.

We next discuss scenario 3. Let N_j^1, N_j^0 denote the numbers of cases and controls, respectively, sampled into study j, $j=2,3$ with $N_T^1 = N_2^1 + N_3^1$ and $N_T^0 = N_2^0 + N_3^0$. The observed-data likelihood is

$$L_3^R(\theta,\psi) = \prod_{i=1}^{N_1} f(\mathscr{D}_i) \prod_{k=1}^{N_2} \int_M f(Y_k|\mathscr{D}_k;\theta,\psi) f(\mathscr{D}_k) dM_k$$
$$\times \prod_{j=1}^{N_3} \left\{\int_E f(Y_j|\mathscr{D}_j;\theta,\psi) f(\mathscr{D}_j) dE_j\right\} P_Y^{-N_T^1} (1-P_Y)^{-N_T^0}.$$

Following the derivations in Section 2.4, the semiparametric log-likelihood is

$$LL_3^R(\theta,\psi) = \sum_{r=1}^{R} n_r \log(p_r) + \sum_{k=1}^{N_2} \log\left\{\sum_r f(Y_k|\mathscr{A}_r;\theta,\psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r\right\}$$
$$+ \sum_{j=1}^{N_3} \log\left\{\sum_r f(Y_j|\mathscr{A}_r;\theta,\psi) I\left[M_j = m_r, \boldsymbol{X}_j = x_r\right] p_r\right\}$$
$$-N_T^1 \log(P_Y) - N_T^0 \log(1-P_Y) \tag{2.19}$$

where P_Y is previously defined. An EM algorithm for maximizing the above expression with respect to $(\gamma,\beta,\boldsymbol{\delta},\mathbf{p})$ is given in the Supplementary Material . Propositions 5 and 6 still hold here. In Supplementary Material, S.4 we provide the asymptotic covariance matrices of $(\gamma,\beta,\boldsymbol{\delta})$ for the two scenarios discussed here.

2.6. Accommodating exposure and mediator interactions

Some approaches for mediation analysis allow for an interaction between the exposure and the mediator (e.g. VanderWeele (2014)). The model in (2.1) can be extended to include an interaction term,

$$\theta = \alpha + \beta M + \gamma E + \omega ME. \tag{2.20}$$

While our approach cannot in general handle an interaction term without incorporating additional information, we can accommodate the special case of a categorical E, where we can assess the effect of M for each exposure group, as described next for the three scenarios in Section 2.1. However, when $\omega \neq 0$ in (2.20), the NIE no longer estimates mediation; but rather the difficult-to-interpret effect of changing the only value of M.

For ease of exposition, we use a binary E and do not adjust for $\mathbf{X}$. We assume that under scenario 1, effect estimates $\boldsymbol{\eta}_{M|E=e} = (\alpha_{M|E=e}, \beta_{M|E=e})$, for $e = 0, 1$, are available. The expectations of the score vectors for $\boldsymbol{\eta}_{M|E=e}$ under the true model (2.20) satisfy

$$\int_{M,E} I(E = e) \left\{ b'(\alpha + \beta M + \gamma E + \omega ME) - b'(\alpha_{M|E=e} + \beta_{M|E=e} M) \right\} dF = 0,$$

$$\int_{M,E} I(E = e) \left\{ b'(\alpha + \beta M + \gamma E + \omega ME) - b'(\alpha_{M|E=e} + \beta_{M|E=e} M) \right\} M dF = 0,$$

for $e = 0, 1$. The above two equations identify $\boldsymbol{\eta} = (\alpha, \beta, \gamma)'$ in model (2.1) and (2.2) without interaction. With estimates from (2.6), $\boldsymbol{\eta} = (\alpha, \beta, \gamma, \omega)$ is identifiable (which is shown similarly to Proposition 1), and $\hat{\boldsymbol{\eta}}$ is the solution of the set of extended score equations given in Supplementary Material, S.5.

Under scenario 2, we assume we have $\boldsymbol{\eta}_{M|E=e} = (\alpha_{M|E=e}, \beta_{M|E=e})$, for $e = 0, 1$, instead of $\boldsymbol{\eta}_M$. Then all parameters can be estimated from the likelihood (2.13) under the constraints $S(\beta)$ and $S(\omega)$ given in Supplementary Material, S.5. Lastly, under scenario 3, we assume that data (Y_j, M_j) for subjects with $E = 0$ and with $E = 1$ are available. The methods in Section 2.4 extend to this scenario by replacing $\int_E f(Y_j|M_j, e; \theta, \psi) f(M_j, e) de$ in the observed likelihood with $f(Y_j|M_j, e = k; \theta, \psi) f(M_j|e = k)$, for $k = 0, 1$. Note that only the conditional probability of M given E has to be modeled. The identifiability and consistency of $\hat{\boldsymbol{\eta}} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\omega})$ are shown by extending Propositions 4–6.

2.7. Estimation of NDE and NIE

We conclude our theoretical derivations by describing the estimation of the NDE_x and NIE_x defined in (2.3) and (2.4), respectively. Under our semiparametric framework, (2.4) corresponds to

$$\mathbb{E}\left[Y\{e, M(e')\}|\boldsymbol{X} = \boldsymbol{x}\right] = \sum_{r=1}^{R} b'(\alpha + \beta m_r + \gamma e + \boldsymbol{\delta}'\boldsymbol{x}) \frac{I(e_r = e', \boldsymbol{x}_r = \boldsymbol{x}) p_r}{\sum_{r=1}^{R} I(e_r = e', \boldsymbol{x}_r = \boldsymbol{x}) p_r},$$

and can be evaluated by plugging the estimated parameters into the above equation. The variance of the estimated NDE and NIE are obtained by applying the delta method or by using numerical simulations based on the asymptotic normal distribution of $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\mathbf{p}})$.

Table 1. Identifiability of causal effects. Extensions of the methods to the settings in this table are provided in Supplementary Material.

Study 1 (M-Y)	**Study 2** (E-Y)	**Causal effects identifiable**
Adjusted for **all** $\boldsymbol{X}$; estimates for $\boldsymbol{X}$ provided	Adjusted for **all** $\boldsymbol{X}$; estimates for $\boldsymbol{X}$ provided (Setting 1)	Yes
	Adjusted for **all** $\boldsymbol{X}$; estimates for $\boldsymbol{X}$ **not provided** (Setting 2)	Yes
Adjusted for $\boldsymbol{X}_1$; estimates for $\boldsymbol{X}_1$ provided	Adjusted for different covariates $\boldsymbol{X}_2$; estimates for $\boldsymbol{X}_2$ provided (Setting 3)	Yes
	Adjusted for $\boldsymbol{X}_2$; estimates for $\boldsymbol{X}_2$ **not provided** (Setting 4)	No; see Remark 3

In scenario 1, where we do not estimate the joint distribution of $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\mathbf{p}})$, we estimate the variance of the estimated NDE and NIE using the following bootstrap approach. At each bootstrap replication, we resample individual-level data $(M, E, \boldsymbol{X})$ with replacement, and simulate new values of $\hat{\boldsymbol{\eta}}_M$ and $\hat{\boldsymbol{\eta}}_E$ from the asymptotic distribution $N(\boldsymbol{\eta}^*, \Sigma_{\boldsymbol{\eta}^*})$. We next estimate $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\mathbf{p}})$ and the NIE and NDE based on this new bootstrap data. Lastly, we estimate the variances of the NDE and NIE using the bootstrap sample variance.

2.8. Remarks on confounder adjustment

Thus far, we have assumed that the models for the associations between M and Y and the association of E and Y are adjusted for the same confounders $\boldsymbol{X}$, and that we have association estimates $\boldsymbol{\delta}$ for $\mathbf{X}$ or individual-level data on $\mathbf{X}$ (setting 1 in Table 1). Supplementary Material provides theoretical justifications and extensions of the methods when $\boldsymbol{\delta}$ is not available or the models are adjusted for different covariates (settings 2 and 3, Table 1).

Remark 3. The NDE and NIE are not identifiable if estimates for some covariates are not provided by either of the two studies. If a data set containing $(Y, M, \boldsymbol{X}_1)$ is available, under some conditions, the regression parameters and causal effects are identifiable even without providing estimates of the effects of the covariates $\boldsymbol{X}_2$ from study 2. Recently Evans et al. (2018) constructed a system of estimating equations that are functions of $\{Y - b'(M, \boldsymbol{X}_1, \boldsymbol{X}_2)\}$ and some vector function of $g(M, \boldsymbol{X}_1)$, with the same dimension as the number of regression coefficients (i.e., $\boldsymbol{\eta}$) in the full model. If the Jacobian of the estimating equations $J = \mathbb{E}\left\{g(M, \boldsymbol{X}_1)\nabla_{\boldsymbol{\eta}} b'(M, \boldsymbol{X}_1, \boldsymbol{X}_2)\right\}$ has full rank, then the regression parameters and causal effects are identifiable. However, such a function may not always exist. For example, if M and $\boldsymbol{X}_1$ are discrete and the number of unique values of $g(M, \boldsymbol{X}_1)$ is smaller than the number of regression coefficients, then J does not have full rank.

2.9. Extension to multiple mediators and/or exposures

Our method can be extended to scenarios with multiple exposures, $\boldsymbol{E} = (E_1, \ldots, E_l)$, multiple mediators, $\boldsymbol{M} = (M_1, \ldots, M_q)$, and covariates, $\mathbf{X}$. These extensions require that the primary data set contains all $\boldsymbol{E}$, $\boldsymbol{M}$, and $\boldsymbol{X}$, and is sufficiently large to model their joint distribution nonparametrically. Supplementary Material, S.6, describes details on a setting in which we have estimated associations between M_j and Y, for $j = 1, \ldots, p$, from p studies, and estimated associations between E_j, for $j = 1, \ldots, q$, and Y from q studies. Using similar steps, one can easily extend our methods to other situations that include information on mediators, covariates and exposures. If some variables are present in multiple studies, the resulting over-determination of the model can be handled in the same way as for the intercept or common covariates, by setting the average of multiple score equations equal to zero.

3. Simulation Studies

We evaluate our procedures for estimating causal effects in several simulation settings. Here, we consider case-control sampling for continuous E and M, and evaluate our procedures for estimating the regression and causal effects for the four settings in Table 1, and for the three scenarios presented in Section 2. Simulations with discrete E and M and continuous Y are described in the Supplementary Material, as well as simulations that assess whether analyzing multiple incomplete data sets with large sample sizes is better than analyzing a small study in which all three variables are measured. Lastly, the Supplementary Material contains simulations that evaluate the effect of omitting an interaction term or covariate adjustment in the model, or using misspecified intercepts α for the binary outcome model.

3.1. Data generation

To estimate the regression and causal effects for the four settings presented in Table 1, we consider case-control sampling for continuous E and M, and two covariates X_1 and X_2. We assume $E \sim N(0,1)$, $X_1 \sim Binom(4, 0.5)$, $X_2 \sim Binom(4, 0.5)$, and $M = E + 1/2X_1 + 1/2X_2 + e$, with $e \sim N(0,1)$. For the first data set we generate $(E_i, M_i, X_{1i}, X_{2i})$, for $i = 1, \ldots, 5000$, from the above model. For the second and third data sets, we generate large cohorts using $logit\{P(Y_i|E_i, M_i)\} = \alpha + \beta E + \gamma M + \delta X_1 + \delta X_2$, where $\alpha = \log(0.01/0.99) = -4.6$, $(\beta, \gamma) = (0.15, 0.15)$, and $\boldsymbol{\delta} \in \{(0,0), (0.1, 0.1)\}$, and sample $N_2^0 = N_3^0 = 1000$ controls and $N_2^1 = N_3^1 = 1000$ cases using a nested design. For each δ, we simulate 10,000 studies. Then we apply our method under the four settings in Table 1. We assume we have a data set with $\mathscr{D} = (E, M, X_1, X_2)$ and the additional information described below.

Setting 1: Both studies are adjusted for X_1 and X_2 and report all estimates. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E, \hat{\delta}_{1E}, \hat{\delta}_{2E})$ and $(\hat{\alpha}_M, \hat{\beta}_M, \hat{\delta}_{1M}, \hat{\delta}_{2M})$, calculated using a logistic regression.

Setting 2: Both studies are adjusted for X_1 and X_2, but one study does not provide estimates. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E, \hat{\delta}_{1E}, \hat{\delta}_{2E})$ calculated using, a logistic regression, and $(\hat{\alpha}_M, \hat{\beta}_M)$ calculated using a logistic regression, adjusting for (X_1, X_2).

Setting 3: Each study is adjusted for one of the two covariates. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E, \hat{\delta}_{1E})$ and $(\hat{\alpha}_M, \hat{\beta}_M, \hat{\delta}_{2M})$, calculated from a logistic regression.

Setting 4: Both studies are adjusted for (X_1, X_2), but do not provide estimates for the covariates. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E)$ and $(\hat{\alpha}_M, \hat{\beta}_M)$, calculated using a logistic regression, adjusting for (X_1, X_2).

To evaluate our procedures for the three scenarios presented in Sections 2.2–2.4, we use similar simulation studies, with the details provided in Section S8 of the Supplementary Material.

Scenario 1: Two summary statistics. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E)$ and $(\hat{\alpha}_M, \hat{\beta}_M)$ calculated using a logistic regression.

Scenario 2: Scenario 2: One summary statistic. We use the marginal estimates $(\hat{\alpha}_E, \hat{\gamma}_E)$ from a logistic regression and a data set with information on (M, Y).

Scenario 3: We have two data sets containing (E, Y) and (M, Y). We also used a traditional mediation analysis, where we have one data set with all three variables as a "gold-standard" (VanderWeele (2014)).

Scenario 0: One complete data set. with (E, M, Y) measured on $N^1 = N^0 = \min(N_2^1, N_3^1)$ cases and controls.

To apply our methods to the three simulated data sets, we disctretize E and M by rounding them to the first digit (i.e., bandwidth of 0.1). The supplementary Material contains results for discretization with a bandwidth of 0.2, starting from zero (e.g., $E = 0.12$ is assigned to $E = 0$; $E = 0.32$ is assigned to $E = 0.2$). We evaluate the mean of the estimated regression coefficients and causal effects, their variances, and the coverage of their 95% confidence intervals (CIs), based on 10,000 simulated studies.

3.2. Results

First, we evaluate the performance of the methods with covariate adjustment. Table 2 shows the means over 10,000 estimated regression coefficients (β, γ),

Table 2. Mean estimates of regression and causal parameters obtained under settings 1–4 in Table 1; Y binary, E and M continuous, (X_1, X_2) discrete.

	$(\gamma, \beta, \delta_1, \delta_2) = (0.15, 0.15, 0, 0)$						$(\gamma, \beta, \delta_1, \delta_2) = (0.15, 0.15, 0.1, 0.1)$					
	$NDE = 0.0196$, $NIE = 0.0179$ at $(X_1, X_2) = (0, 1)$						$NDE = 0.021$; $NIE = 0.0189$ at $(X_1, X_2) = (0, 1)$					
Setting	β	γ	δ_1	δ_2	NDE	NIE	β	γ	δ_1	δ_2	NDE	NIE
1	0.147	0.152	-0.001	-0.001	0.0190	0.0183	0.146	0.153	0.098	0.099	0.020	0.0199
2	0.147	0.151	-0.001	0.001	0.0191	0.0182	0.146	0.152	0.099	0.098	0.020	0.0197
3	0.149	0.152	-0.005	0.001	0.0192	0.0183	0.142	0.152	0.096	0.097	0.0195	0.020
4	0.121	0.177	NA	NA	0.015	0.020	0.120	0.178	NA	NA	0.0150	0.020

NDEs, and NIEs under the four settings in Table 1. All three methods for settings 1, 2, and 3 produce unbiased estimates of β, γ, and NDE and NIE. As expected, when no association estimates for the covariates are provided, the estimates of (β, γ), NDE, and NIE are biased. This highlights the importance of incorporating covariate estimates in our methods.

Next, we summarize the results from the four sets of simulations that evaluate the methods for the three scenarios in Section 2.1. The mean of the 10,000 estimated regression coefficients (β, γ), the NDE, and the NIE are virtually unbiased for scenarios 1 and 2 and all parameter values ("mean" in Table 3). The mean of the 10,000 analytical standard errors ("SE" in Table 3) agrees well with the empirical standard errors and is similar for the same three scenarios. The 95% Wald-based CIs have close to nominal coverage ("cov" in Table 3). In scenario 3, the estimates of the regression coefficients, NDE, and NIE are slightly biased, and the coverage of the 95% CIs is somewhat lower than 95%. The bias disappeares and the CI coverage improves as the sample size increases (Supplementary Material, Table S1) or we use a bandwidth for discretization of 0.2 (Supplementary Tables S2 and S3). Under scenario 0 (inference using one data set with all variables), the standard deviations of the regression coefficients, NDE, and NIE are noticeably smaller.

The Supplementary Material, Tables S4–S9 show additional results. When E and M are discrete with either continuous or binary Y, the estimated regression coefficients, NDEs, and NIEs are virtually unbiased for all settings and parameter values (Supplementary Material, Tables S4–S6). The mean of the standard errors agrees well with the empirical standard errors, and the coverage of the 95% Wald-based CIs is close to nominal. For continuous E, M, and Y (Supplementary Material, Table S7), the estimated regression coefficients, NIE, and NDE in scenarios 1 and 2 are unbiased, and the analytical variances agree well with the empirical variances. The 95% CIs have close to nominal coverage. In scenario 3, the estimated regression coefficients, NDE and NIE are slightly biased, and the covareage of 95% CI is somewhat too low. However, similarly to case of binary Y, the bias decreases and the CI coverage improves with increasing sample size (Supplementary Material, Table S1) or when the bandwidth of discretization

Table 3. Simulation results for scenarios with binary Y, and continuous E and M. Mean: average value of an estimator. SE: average value of the analytical standard error of the estimator. Cov: coverage of the true value by the 95% confidence interval computed using the asymptotic variance.

	β		γ		NDE		NIE	
	Mean (SE)	Cov	Mean (SE)	Cov	Mean (SE)	Cov	Mean (SE)	Cov
$\gamma = 0.15$; $\beta = 0.15$; $NDE = 0.016$; $NIE = 0.015$								
Scenario 0:	0.15 (0.05)	0.95	0.15 (0.06)	0.95	0.016 (0.007)	0.96	0.015 (0.006)	0.94
Scenario 1:	0.15 (0.08)	0.95	0.15 (0.11)	0.95	0.016 (0.012)	0.95	0.015 (0.010)	0.92
Scenario 2:	0.15 (0.07)	0.95	0.14 (0.10)	0.95	0.015 (0.011)	0.95	0.015 (0.009)	0.92
Scenario 3:	0.14 (0.06)	0.94	0.15 (0.09)	0.94	0.016 (0.009)	0.94	0.014 (0.008)	0.87
$\gamma = 0.15$; $\beta = 0$; $NDE = 0.014$; $NIE = 0$								
Scenario 0:	0 (0.04)	0.96	0.15 (0.06)	0.95	0.014 (0.006)	0.95	0 (0.004)	0.97
Scenario 1:	0 (0.11)	0.95	0.15 (0.08)	0.95	0.015 (0.010)	0.94	-0.001 (0.012)	0.95
Scenario 2:	0 (0.10)	0.95	0.15 (0.07)	0.95	0.015 (0.009)	0.94	0 (0.010)	0.95
Scenario 3:	0.02 (0.09)	0.93	0.13 (0.06)	0.92	0.013 (0.007)	0.90	0.002 (0.009)	0.93
$\gamma = 0$; $\beta = 0.15$; $NDE = 0$; $NIE = 0.015$								
Scenario 0:	0.15 (0.04)	0.95	0 (0.06)	0.95	0 (0.007)	0.96	0.015 (0.006)	0.94
Scenario 1:	0.15 (0.11)	0.95	0 (0.08)	0.95	0 (0.008)	0.98	0.014 (0.011)	0.95
Scenario 2:	0.14 (0.10)	0.95	0.01 (0.07)	0.95	0.001 (0.007)	0.98	0.013 (0.010)	0.95
Scenario 3:	0.13 (0.09)	0.93	0.01 (0.06)	0.93	0.001 (0.006)	0.97	0.012 (0.008)	0.93
$\gamma = 0$; $\beta = 0$; $NDE = 0$; $NIE = 0$								
Scenario 0:	0 (0.04)	0.95	0 (0.06)	0.96	0 (0.006)	0.96	0 (0.004)	0.97
Scenario 1:	0 (0.11)	0.95	0 (0.08)	0.95	0 (0.008)	0.98	0 (0.010)	0.95
Scenario 2:	0 (0.10)	0.95	0 (0.07)	0.95	0 (0.007)	0.98	0 (0.009)	0.95
Scenario 3:	0 (0.09)	0.94	0 (0.06)	0.94	0 (0.006)	0.97	0 (0.008)	0.94

increases to 0.2 (Supplementary Material, Tables S8 and S9).

We examine the effect of changing the sample size for one of the three studies on the variance of the NDEs and NIEs (Supplementary Material, Figures S1, S2). Increasing N_1, the sample size of the study with data on (E, M), did not improve the efficiency or reduce the variance in any of the three scenarios (Supplementary Figures S1A and S2A). Increasing N_2, the sample size of the study with measurements on (E, Y), reduces the variances of the NDE and NIE (Supplementary Material, Figures S1B and S2B). However, even for $N_2 = 10000$, the variances of the estimates are usually larger than those of a small study that measures all three variables. Increasing N_3, the sample size of the study with (M, Y), results in lower variances for both the NDE and the NIE (Supplementary Material, Figures S1C and S2C), but the estimates are again less precise than those of a small study with all three variables.

Lastly, we investigate the effects of a model misspecification on the estimates of the regression coefficients and the causal effects (Supplementary Material, Figures S3-S6). Omitting the interaction term δ from the model results in biased estimates of the NDE that linearly increases with δ (Supplementary Material,

Figures S3 and S4). For a binary Y, the bias is also observed in the estimates of the regression coefficients. Omitting confounders for the relationship between M and E leads to bias in both regression coefficients, and as a result, in the NDE and NIE. These results again highlight the importance of confounder adjustments. The results in Table S10 demonstrate that when Y given E and M has a logistic distribution and in the populations $P(Y = 1) << 1$, misspecifing $\alpha = \alpha_E = \alpha_M$ does not noticeably bias $\boldsymbol{\eta}$.

4. Data Example

Genome-wide association studies (GWAS) have identified hundreds of genetic variants that affect a woman's breast cancer risk. Individual variants have only weak associations with breast cancer risk, but when combined, the resulting polygenic risk score (PRS) is strongly associated. The remaining question is how this PRS and the underlying genetics affect the risk of breast cancer. In an attempt to answer this question, studies have evaluated the relationship between the PRS and various breast cancer risk factors, identifying an association with the number of terminal duct lobular units (TDLUs), milk-producing breast structures known to be associated with breast cancer risk.

We use our new approach to determine the proportion of the effect of PRS (E) on breast cancer status (Y) explained by the TDLU count (M). We have (i) a data set with PRS scores and the TDLU counts for 1,398 women (Bodelon et al. (2020)), (ii) a 4×2 table with the numbers of breast cancer cases and controls by TDLU count categories (quartiles) in a case-control study (Figueroa et al. (2014)), and (iii) an OR and 95% CI describing the effect of a one-standard deviation (1-SD) increase in PRS on breast cancer risk (Mavaddat et al. (2019)). Details onf the study-specific results are given in Supplementary Material, Figure 7. We use the method for two marginal estimates described in Section 2.2 and assume the five-year risk of breast cancer is 2% (Mavaddat et al. (2019)). We discretize the PRS into bins of length 0.1.

The overall PRS effect on breast cancer risk is summarized by an OR = 1.65 (95% nCI: 1.59 – 1.72). Conditioned on the TDLU count, we estimate the conditional OR = 1.63 (1.56 – 1.71), suggesting that the TDLU count does not account for a significant proportion of the overall effect. We further estimated the NIE and NDE. Assuming that the overall breast cancer risk was 2% in the population, a 1-SD increase in PRS directly increases breast cancer risk by 0.44% (0.31 – 0.58), and indirectly by a nonsignificant 0.074% (-0.1 – 0.28). The results are similar for sensitivity analyses when we vary the marginal effect of the TDLU count on breast cancer risk to assess possible differences between the populations of the two breast cancer studies.

5. Discussion

We have proposed novel semiparametric approaches for mediation analysis to estimate the NDE and NIE under three scenarios that arise in practice when the exposure, mediator, and outcome are not measured in a single study. We have demonstrated that all regression parameters are identifiable under the generalized linear model (2.1 and 2.2), and estimates are consistent and asymptotically normal. We discussed an extension to allow for interactions between the mediator and the exposure in some settings, and how to accommodate multiple mediators, multiple exposures, and confounders.

We highlight the key features of our method. Most importantly, simulation studies with small sample sizes show that our approaches yield unbiased estimates and confidence intervals with nominal coverage. Moreover, for continuous outcomes, the estimates obtained when only summary statistics were available (scenario 1) were as efficient as the estimates obtained when there were partially observed data (scenarios 2, 3). For binary outcomes, the estimates from scenario 1 were less efficient than those of the other two scenarios, and the estimates of the regression coefficients, NDE, and NIE from scenario 3 were about 20% more efficient. However, the efficiency of the estimates from our methods is noticeably lower than that of estimates based on a single data set that contains all relevant variables.

Our approach builds on existing statistical methods that combine information from multiple data sets to estimate parameters. However, these methods, discussed below, require that at least one data set contains all relevant variables. First, our research is closely related to related to methods for two-phase and outcome-dependent sampling designs, where a subset of units is selected from a large data set to measure additional variables of interest (e.g., Lin and Tang (2011)). Methods based on semiparametric maximum likelihood that account for a sampling design have been proposed to analyze such studies (e.g., Breslow and Holubkov (1997); Lin and Zeng (2006)). Several approaches, including methods based on calibration equations (Chen and Chen (2000)), regression imputation (Cheng et al. (2019)), and inverse probability weighting (Cao, Tsiatis and Davidian (2009)), have been proposed to combine a small study with a complete set of variables with a large external data set with fewer variables in order to improve the regression coefficient efficiency. Many of these methods require access to individual-level data from both data sets. Chatterjee et al. (2016) and Zhang et al. (2020) proposed a constrained MLE for model calibration using summary-level information from multiple sources. Second, our model assumes that the three sources of information are based on samples from the same underlying population. Many methods have been developed to extend causal inferences from a one population to another population under generalizability or transportability assumptions (see ,e.g., Buchanan et al. (2018)). Under these

assumptions, Yang and Ding (2020) and Evans et al. (2018) proposed approaches for combining multiple data sets in order to estimate the causal effects of an exposure on an outcome that can handle nonrepresentative sampling. If we had a common set of covariates across all our studies, we could adapt these methods to our approach, but this extension is beyond the scope of this study.

We highlight some limitations of the proposed framework. First, the model specified in Section 2.1 assumes no interaction between the exposure and the mediator. In the presence of an interaction, the estimates of NIE and NDE are biased. If no additional data with all three measurements are available, then including the unknown interaction parameter in the outcome model in Section 2.1 causes identifiability problems. Thus, in sensitivity analyses to assess the potential bias in NIE and NDE, one can model the interaction term as a linear function of γ, $\omega = k\gamma$, where k represents a specified proportion of the additive effect of the exposure. Alternatively, when individual-level data on $(Y, E, \mathbf{X})$ or $(Y, M, \mathbf{X})$ are available, we can estimate the interaction between E and M by adapting the methods of Evans et al. (2018).

Second, for scenarios 2 and 3 ,the likelihood functions involve estimating the joint density of $(E, M, \boldsymbol{X})$, which is challenging for continuous variables bacause our method requires all unique values of $(E, \boldsymbol{X})$ or $(M, \boldsymbol{X})$ observed in all individual-level data sets. Our approach of discretizing continuous variables can produce biased results and a loss of efficiency. One solution is to build on ideas for handling expensive continuous variables in two-phase studies (e.g., Zeng and Lin (2014)). In future work, we plan to use kernel functions to model the joint distribution of $(E, M, \mathbf{X})$, as outlined for scenarios one and two. Despite these limitations, our proposed methods are practically important novel tools for mediation analysis with partially observed data.

Supplementary Material

The online Supplementary Material contains appendices, tables and figures referenced in Sections 2, 3.2, and 4.

Acknowledgments

This study used the computational resources of the NIH HPC Biowulf cluster. We thank Clara Bodelon for access to the data used in our example. This study was funded by the NIH/NCI Cancer Center Support grant P30 CA008748.

References

Allendoerfer, C. B. (1974). *Calculus of Several Variables and Differentiable Manifolds.* Macmillan.

Bodelon, C., Oh, H., Derkach, A., Sampson, J. N., Sprague, B., Vacek, P. et al. (2020). Polygenic risk score for the prediction of breast cancer is related to lesser terminal duct lobular unit involution of the breast. *NPJ Breast Cancer* **6**, 1–6.

Breslow, N. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 447–461.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S. et al. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181**, 1193–1209.

Cao, W., Tsiatis, A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Carroll, R., Wang, S. and Wang, C. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* **90**, 157–169.

Chatterjee, N., Chen, Y., Maas, P. and Carroll, R. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**, 107–117.

Chen, Y. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 449–460.

Cheng, J., Cheng, N. F., Guo, Z., Gregorich, S., Ismail, A. I. and Gansky, S. A. (2018). Mediation analysis for count and zero-inflated count data. *Statistical methods in medical research* **27**, 2756–2774.

Cheng, W., Taylor, J. G., Gu, T., Tomlins, S. and Mukherjee, B. (2019). Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68**, 121–139.

Daniels, M. J., Roy, J., Kim, C., Hogan, J. and Perri, M. (2012). Bayesian inference for the causal effect of mediation. *Biometrics* **68**, 1028–1036.

Derkach, A., Pfeiffer, R., Chen, T. and Sampson, J. (2019). High dimensional mediation analysis with latent variables. *Biometrics* **75**, 745–756.

Evans, K., Sun, B., Robins, J. and Tchetgen, E. J. T. (2018). Doubly robust regression analysis for data fusion. *arXiv:1808.07309*.

Figueroa, J. D., Pfeiffer, R. M., Patel, D. A., Linville, L., Brinton, L. A., Gierach, G. L. et al. (2014). Terminal duct lobular unit involution of the normal breast: Implications for breast cancer etiology. *JNCI: Journal of the National Cancer Institute* **106**, dju286.

Huang, Y.-T. (2019). Variance component tests of multivariate mediation effects under composite null hypotheses. *Biometrics* **75**, 1191–1204.

Huang, Y.-T. and Cai, T. (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics* **72**, 563–574.

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

Imai, K., Keele, L. and Yamamoto, T. (2011). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25**, 51–71.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–585.

Lin, D. and Tang, Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics* **89**, 354–367.

Lin, D. Y. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association* **101**, 89–104.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A. et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics* **104**, 21–34.

VanderWeele, T. J. (2014). A unification of mediation and interaction: A four-way decomposition. *Epidemiology* **25**, 749–761.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association* **115**, 1540–1554.

Yuan, K. and Jennrich, R. (2000). Estimating equations with nuisance parameters: Theory and applications. *Annals of the Institute of Statistical Mathematics* **52**, 343–350.

Zeng, D. and Lin, D. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association* **109**, 371–383.

Zeng, S., Rosenbaum, S., Alberts, S. C., Archie, E. A. and Li, F. (2021). Causal mediation analysis for sparse and irregular longitudinal data. *The Annals of Applied Statistics* **15**, 747–767.

Zhang, H., Deng, L., Schiffman, M., Qin, J. and Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107**, 689–703.

Andriy Derkach

Department of Epidemiology and Biostatistics, MSKCC, New York, NY 10017, USA.

E-mail: derkacha@mskcc.org

Joshua N. Sampson

Biostatistics Branch, DCEG, NCI, NIH, Rockville, MD 20852, USA.

E-mail: joshua.sampson@nih.gov

Ruth M. Pfeiffer

Biostatistics Branch, DCEG, NCI, NIH, Rockville, MD 20852, USA.

E-mail: pfeiffer@mail.nih.gov

(Received October 2021; accepted October 2022)