Statistica Sinica Preprint No: SS-2021-0373	
Title	Integrating Incomplete Data for Mediation Analysis
Manuscript ID	SS-2021-0373
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0373
Complete List of Authors	Andriy Derkach, Joshua N. Sampson and Ruth M. Pfeiffer
Corresponding Authors	Andriy Derkach
E-mails	derkacha@mskcc.org
Notice: Accepted version subject to English editing.	

Statistica Sinica

Integrating incomplete data for mediation analysis

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

[1]*Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center*

[2]*Biostatistics Branch, DCEG, National Cancer Institute*

Abstract: Mediation analysis examines the relationships between an exposure, a mediator, and an outcome. There are many approaches available for performing mediation analysis, but they all require access to a single complete dataset that contains the three key variables: outcome, exposure, and mediator. Here, we propose semi-parametric methods for mediation analysis to estimate the standard causal parameters (direct and indirect effects) by combining information from several incomplete datasets, each containing only two of the three key variables. Importantly, our methods also handle scenarios when only summary statistics based on those datasets 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 in simulations and quantify the loss in efficiency from lack of a complete dataset with all three variables. We then determine if 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; semi-parametric likelihood; summary level information.

1. Introduction

Mediation analysis has become a very 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 nimbly 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 dataset. Here, we develop methods for mediation analysis when we have three "incomplete" datasets, each containing only two of the 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 mediated by the number of terminal duct lobular units (TDLUs, M). TDLU are epithelial structures that produce milk during lactation. We measured the PRS and number of TDLUs in a cohort of women and therefore had a dataset 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) were available. Thus, we aimed to perform mediation analysis by combining information from these three incomplete datasets.

Here we develop methods to perform mediation analysis in three scenarios. For all scenarios we assume we have a dataset containing individual-level data on E and M. In the first scenario we have two summary statistics (e.g. published ORs) that capture the relationship between E and Y and the relationship between M and Y. In the second scenario we have a summary statistic describing one relationship (either between E and Y, or M and Y) and a dataset with individual-level measurements for evaluating the second relationship. In the third scenario we have two additional datasets, 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 proposed 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 difference 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 be also extended for 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 upon methods that use semi-parametric 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 constrained maximum likelihood estimation (Chatterjee et al., 2016; Zhang et al., 2020; Kundu et al., 2019). However, all these methods assume that at least one dataset 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 study the estimates in finite samples in simulations (Section 3) and analyze the motivating breast cancer data (Section 4) before a discussion in Section 5.

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,

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

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 the natural direct effect (NDE) and the natural indirect effect (NIE), while controlling for $\boldsymbol{X}$,

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

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), (2.2) and assumptions outlined in Imai et al. (2010, 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, semi-parametric and non-parametric approaches for estimating 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 datasets. Specifically, we consider three scenarios that are defined by the available information. For all scenarios, we assume that we have a dataset with individual level information on $(M, E, \boldsymbol{X})$ and all three sources of information are based on samples from the same underlying population. The three scenarios are

1. We have estimates of the associations between M and Y and between E and Y with both models 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 dataset with individual-level data on $(M, E, \boldsymbol{X})$, or, alternatively, we have an estimate of the association of E and Y adjusted for $\boldsymbol{X}$ and individual-level data on $(Y, M, \boldsymbol{X})$.

3. We have two datasets, 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 where the datasets have different sets of covariates in 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 dataset with individual-level measurements $\mathscr{D}_i = (M_i, E_i, \mathbf{X}_i)$, $i = 1, \ldots, N_1$. We also have an estimate of the marginal association between M and Y adjusted for $\mathbf{X}$. In other words, there was a prior study that collected data on $(M, Y, \mathbf{X})$, assumed

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

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, there was another study that collected data on $(E, Y, \mathbf{X})$, assumed

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

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 datasets 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 $\boldsymbol{\eta}_E = (\alpha_E, \gamma_E, \boldsymbol{\delta}_E)$. We now propose a new semi-parametric 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 $\mathscr{D}_i,\ i = 1, \ldots, N_1$.

Letting ∇ denote the gradient operator that yields the vector of partial derivatives, the score vectors for the working models (i.e. the models that does 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 $\boldsymbol{\eta}$,

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

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

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 do assume that the joint distribution of $(Y, M, E, \boldsymbol{X})$ is the same in the two studies.

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

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

We similarly obtain three equations based on (2.8). The intercept α and the covariate specific parameters $\boldsymbol{\delta}$ are present in the equations (2.7) based on $\boldsymbol{U}_M$ and (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

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

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. 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 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 already points to the proof of the proposition for outcomes Y with distributions in the exponential family that 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 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 are applying the inverse function theorem (Allendoerfer, 1974, "Theorems about Differentiable Functions"). The inverse function theorem states that the corresponding objective function to system of equations (2.9) will be strictly convext with a unique minimum iff the Jacobian with respect to 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

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

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 assumed that the intercepts α_M, α_E are provided and the true intercept α is obtained based on 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 $\boldsymbol{\delta}$ ignoring the intercept. Let $\mathbb{E}(M) = \mathbb{E}(E) = \mathbb{E}(X_1) = \ldots = \mathbb{E}(X_k) = 0$. 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 β, γ and δ_j, $j = 1, \ldots, k$, 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 a cases-control study do not follow the population distribution of $\mathscr{D}_i$ (Carroll et al., 1995). However, α is not estimable and we propose to set $\alpha = \log\{P(Y = 1)/P(Y = 0)\}$ when β, γ and $P(Y = 1)$ are small. In

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

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

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 $N_i \to \infty, i = 1, 2, 3$ 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 the equations (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 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

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

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 dataset and a marginal estimate: scenario 2

We now assume that in addition to the study with individual-level data on $\mathscr{D}_i,\ i = 1, \ldots, N_1$, we have another (incomplete) dataset with individual-level data on $(Y_j, E_j, \boldsymbol{X}_j),\ j = 1, \ldots, N_2$ 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.

2.3 Estimation based on one incomplete dataset and a marginal estimate: scenario 2

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

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

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

under the constraint $\sum_{r=1}^{R}\left\{E(Y|\mathscr{d}_r;\boldsymbol{\eta})-E(Y|m_r,\boldsymbol{x}_r;\boldsymbol{\eta}_M)\right\}\mathbf{V}$, 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

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

In the Appendix we propose a computationally efficient and numerically robust expectation and maximization (EM) algorithm for maximizing expression (2.14) as a function of $\boldsymbol{\eta}$ and $\boldsymbol{p}$. Next, we demonstrate uniqueness of the maximum and establish 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 $N_1 \to \infty$, $N_k/N_1 \to \rho_k$ (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 additionally

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

where J_η is the asymptotic covariance matrix of $\boldsymbol{\eta}$ defined in Appendix S.3.

The proofs for these propositions are given in Appendices S.2 and S.3.

2.4 Estimation based on two incomplete datasets: scenario 3

Here we assume that three incomplete datasets with individual level data are provided, a study with measures $\mathscr{D}_i$, $i = 1, \ldots, N_1$, a study with measures $(Y_k, E_k, \boldsymbol{X}_k)$, $k = 1, \ldots, N_2$ and a study with measures $(Y_j, M_j, \boldsymbol{X}_j)$

for $j = 1, \ldots, N_3$. The observed-data likelihood is

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

Again, $f(\mathscr{D})$ cannot be factored out or estimated only from the first dataset without 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 dataset with measurements on $\mathscr{D}$, and estimate F non-parametrically 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

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

In the Appendix 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.

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 Appendix S.3.

Our semi-parametric 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 in Appendices S.2 and S.4.

2.5 Estimation under case-control sampling

So far we assumed that the distributions of $(Y, E, M, \boldsymbol{X})$ in the three data sources are the same. We now consider the setting when 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 when one study is based on case-control sampling is straight forward. As 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 to use $\alpha \approx \alpha_E \approx \alpha_M$ 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 thus 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 sampled into the second study, respectively, and $P_Y = P(Y=1) = \int_{\mathscr{D}} P(Y=1|\mathscr{D};\boldsymbol{\eta})dF(\mathscr{D})$ 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 semi-parametric log-likelihood is

$$LL_2^R(\theta,\psi) = \sum_{r=1}^{R} n_r \log(p_r) + \sum_{k=1}^{N_2} \log \left\{ \sum_r f(Y_k|\boldsymbol{\mathscr{d}}_r;\theta,\psi) I\left[E_k = e_r, \boldsymbol{X}_k = \boldsymbol{x}_r\right] p_r \right\}$$

$$-N_2^1 \log(P_Y) - N_2^0 \log(1-P_Y) + \boldsymbol{\lambda}' N_2 \sum_r \{E(Y|\boldsymbol{\mathscr{d}}_r;\boldsymbol{\eta}) - E(Y|m_r,\boldsymbol{x}_r;\boldsymbol{\eta}_M)\} \begin{pmatrix} m_r \\ \boldsymbol{x}_r \end{pmatrix} p_r,$$

where $P_Y = \sum_r P(Y=1|\boldsymbol{\mathscr{d}}_r,\boldsymbol{x}_r;\theta) p_r$ and $\boldsymbol{\lambda}$ is the vector of Lagrange multiplier. Appendix 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 semi-parametric log-likelihood is

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

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 Appendix. Propositions 5 and 6 still hold here. In Appendix 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,

$$\theta = \alpha + \beta M + \gamma E + \omega ME. \tag{2.18}$$

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 (2.18), the NIE no longer estimates mediation but the difficult-to-interpret effect of only changing the 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 $\boldsymbol{\eta}_{M|E=e} = (\alpha_{M|E=e}, \beta_{M|E=e}), e = 0, 1$, are available. The expectations of the score vectors for $\boldsymbol{\eta}_{M|E=e}$ under the true model (2.18) satisfy

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

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

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 Appendix S.5.

Under scenario 2, we assume we have $\boldsymbol{\eta}_{M|E=e} = (\alpha_{M|E=e}, \beta_{M|E=e}), e = 0, 1$ 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 Appendix 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 by $f(Y_j|M_j, e = k; \theta, \psi) f(M_j|e = k), k = 0, 1$. Note that only the conditional

probability of M given E has to be modeled. 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). 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 numerical simulations based on the asymptotic normal distribution of $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\boldsymbol{\delta}}, \hat{\mathbf{p}})$.

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 individaul-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

So far we assumed that the models for the associations between M and Y and the association of E and Y were adjusted for the same confounders $\boldsymbol{X}$, and the association estimates $\boldsymbol{\delta}$ for $\mathbf{X}$ or individual level data on $\mathbf{X}$ are provided (setting 1 in Table 1). Supplemental 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).

Table 1: **Identifiability of causal effects.** Details for extensions of the methods to the settings in this table are provided in Supplemental 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

Remark 3. The NDE and NIE are not identifiable if estimates for some covariates are not provided by any of the two studies. If a dataset contain-

ing $(Y, M, \boldsymbol{X}_1)$ is available, under some conditons, the regression parameters and causal effects are indentifiable even without providing estimates of the effects of 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 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, $\mathbf{X}$. This extension requires that the primary dataset contains all $\boldsymbol{E}$, $\boldsymbol{M}$, and $\boldsymbol{X}$ and is sufficiently large to model their joint distribution non-parametrically. Appendix S.6 gives details on a setting where 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 are provided. Using

similar steps one can easily extend our methods to other situations where studies contain information on mediators, covariates and exposures. If some variables are present in multiple studies, the resulting over-determination of the model can be handled as for the intercept or common covariates, by setting the average of multiple score equations equal to zero.

3. Simulation studies

We evaluated our procedures for estimating causal effects in several simulation settings. Here, we considered case-control sampling for continuous E and M and evaluated our procedures for estimating regression and causal effects for the four settings in Table 1, and for the three scenarios presented in Section 2. Simulations with discrete E, M and continuous Y are described in Supplemental Material. Supplemental Material also includes simulations to assess if analyzing multiple incomplete datasets with large sample sizes is better than analyzing a small study in which all three variables are measured. Lastly, the Supplemental Material contains simulations that evaluate the impact of omitting an interaction term or covariate adjustment in the model or using mis-specified intercepts α for the binary outcome model.

3.1 Data generation

For estimating regression and causal effects for the four settings presented in Table 1, we considered case-control sampling for continuous E, M and two covariates X_1 and X_2. We assumed $E \sim N(0,1)$, $X_1 \sim Binom(4, 0.5)$, $X_2 \sim Binom(4, 0.5)$, and $M = E + 1/2X_1 + 1/2X_2 + e$, with $e \sim N(0,1)$. For the first dataset we generated $(E_i, M_i, X_{1i}, X_{2i}), i = 1, \ldots, 5000$ from the above model. For the second and third datasets, we generated large cohorts with $logit\{P(Y_i|E_i, M_i)\} = \alpha + \beta E + \gamma M + \delta X_1 + \delta X_2$, 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 sampled $N_2^0 = N_3^0 = 1000$ controls and $N_2^1 = N_3^1 = 1000$ cases using a nested design. For each δ we simulated 10,000 studies. Then we applied our method under the four settings in Table 1. We assume we have a dataset with $\mathscr{D} = (E, M, X_1, X_2)$ and the additional information described below.

Setting 1: Both studies adjusted for X_1 and X_2 and report all estimates. We utilize 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 logistic regression.

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

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

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

To evaluate our procedures for the three scenarios presented in Sections 2.2-2.4, we used similar simulation studies with details shown in Section S8, Supplemental Materials.

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

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

Scenario 3: We have two datasets containing (E, Y) and (M, Y).

We also used traditional mediation analysis where we have one dataset with all three variables as a "gold-standard" (VanderWeele, 2014).

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

To apply our methods we discretized E and M by rounding them to the first digit in three simulated datasets (i.e bandwidth of 0.1). Supplemental

material contains results for discretization with a bandwidth of 0.2 starting from 0 (e.g. $E = 0.12$ is assigned to $E = 0$; $E = 0.32$ is assigned to $E = 0.2$). We evaluated 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 evaluated the performance of the methods with covariate adjustment. Table 2 shows means over 10,000 estimated regression coefficients (β, γ), NDEs and NIEs under the four settings in Table 1. All three methods for settings 1, 2 and 3 produced unbiased estimates of β, γ and NDE and NIE. As expected, when association estimates for covariates were not provided, estimates of (β, γ), NDE and NIE were biased. This highlights the importance of incorporating covariate estimates into 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 10,000 estimated regression coefficients (β, γ), NDE's and NIE's were 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) agreed well with the empirical standard errors and was similar for

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

the same three scenarios. The 95% Wald based CIs had close to nominal coverage ("cov" in Table 3). In scenario 3, the estimates of the regression coefficients, NDE and NIE were slightly biased, and the coverage of the 95% CIs was somewhat lower than 95%. The bias disappeared and the CI coverage improved with increasing sample size (Supplemental Table S1) or used a bandwidth for discretization of 0.2 (Supplemental Tables S2 and S3). Under scenario 0 (inference using one dataset with all variables), the standard deviations of the regression coefficients, NDE and NIE were noticeably smaller.

Supplemental Tables S4-S9 shows additional results. When E and M were discrete with either continuous or binary Y, the estimated regression coefficients, NDEs and NIEs were virtually unbiased for all settings and

Table 3: Simulation results for scenarios with binary Y, 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 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

parameter values (Supplemental Tables S4-S6). The mean of the standard errors agreed well with the empirical standard errors and the coverage of the 95% Wald based CIs was close to nominal. For continuous E, M and Y (Supplemental Table S7), the estimated regression coefficients, NIE, and NDE in scenarios 1 and 2 were unbiased, and the analytical variances agreed well with the empirical variances. The 95% CIs had close to nominal coverage. In scenario 3, the estimated regression coefficients, NDE and NIE were slightly biased, and the 95% CI coverage was somewhat too low. However, similarly to binary Y, the bias decreased and the CI coverage improved with increasing sample size (Supplemental Table S1) or when the bandwidth of discretization was increased to 0.2 (Supplemental Tables S8 and S9).

We studied the impact of changing the sample size of one of the three studies on the variance of the NDEs and NIEs (Supplemental Figures S1, S2). Increasing N_1, the sample size of the study with data on (E, M), did not improve efficiency or reduce the variance in any of the three scenarios (Supplemental Figures S1A and S2A). Increasing N_2, the sample size of the study with measurements on (E, Y), reduced the variances of the NDE and NIE (Supplemental Figures S1B and S2B). However, even for $N_2 = 10{,}000$ the variances of the estimates were usually larger than for a small study that measured all three variables. Increasing N_3, the sample size of the

study with (M, Y), resulted in lower variances of both the NDE and NIE (Supplemental Figures S1C and S2C), but estimates were again less precise than for a small study with all three variables.

Lastly, we investigated the effects of model misspecification on estimates of the regression coefficients and causal effects (Supplemental Figures S3-S6). Omitting the interaction term δ from the model resulted in biased estimates of NDE that linearly increased with δ (Supplemental Figures S3 and S4). For binary Y, the bias was also observed in the estimates of the regression coefficients. Omitting confounders for the relationship between M and E led 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 $P(Y = 1) << 1$, then misspecifing $\alpha = \alpha_E = \alpha_M$ does not noticeably bias $\boldsymbol{\eta}$.

4. Data example

Genome-Wide Association Studies (GWAS) have identified 100's 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 outstanding

question is how this PRS and the underlying genetics affect the risk of breast cancer. To answer it, studies have evaluated the relationship between this PRS and various breast cancer risk factors, and identified an association with the number of terminal duct lobular units (TDLUs), milk-producing breast structures known to be associated with breast cancer risk.

We used our new approach to determine the proportion of the effect of PRS (E) on breast cancer status (Y) explained by TDLU count (M). We had (i) a dataset with PRS scores and TDLU counts for 1398 women (Bodelon et al. (2020)), (ii) a 4 x 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)), (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 on study-specific results are given in Supplementary Figure 7. We used the method for two marginal estimates described in Section 2.2 and assumed the 5 year-risk of breast cancer was 2% (Mavaddat et al., 2019). We discretized the PRS into bins of length 0.1.

The overall PRS effect on breast cancer risk was summarized by an OR = 1.65 (95%CI: 1.59 - 1.72). Conditioned on TDLU count, we estimated the conditional OR = 1.63 (1.56 - 1.71), suggesting that TDLU count did 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 increased breast cancer risk by 0.44% (0.31 - 0.58) and indirectly by a non-significant 0.074% (-0.1 - 0.28). Results were similar in sensitivity analyses when the marginal effect of TDLU on breast cancer risk was varied to assess possible differences between the populations of the two breast cancer studies.

5. Discussion

We proposed novel semi-parametric approaches for mediation analysis to estimate NDE and NIE under three scenarios that arise in practice when the exposure, mediator, and outcome are not measured in a single study. We 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 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 showed that our approaches yielded 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 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 was noticeably lower than that of estimates based on a single dataset containing all relevant variables.

Our approach builds upon previously published statistical methods that combine information from multiple datasets to estimate parameters. However, these methods, discussed below, require that at least one dataset 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 dataset to measure additional variables of interest (e.g. Lin and Tang, 2011). Methods based on semiparametric maximum likelihood accounting 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 et al., 2009) have been proposed to combine a small study with a complete set of variables with a

large external dataset with fewer variables to improve the regression coefficients' efficiency. Many of these methods require access to individual-level data from both datasets. More recently Chatterjee et al. (2016) and Zhang et al. (2020) proposed constrained maximum likelihood estimation for model calibration using summary-level information from multiple sources. Second, our model assumes that 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 recently, Yang and Ding (2020) and Evans et al. (2018) proposed approaches for combing multiple datasets to estimate causal effects of an exposure on outcome that can handle non-representative sampling. If we had a common set of covariates across all our studies, we could adapt these methods to our approach, but this extension goes beyond the scope of the current paper.

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 interaction, estimates of NIE and NDE are biased. If no additional data with all three measurements are available, including the unknown interaction parameter in the outcome model in Sec-

tion 2.1 causes identifiability problems. Thus, in sensitivity analyses to assess 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 could 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 as our method require all unique values of $(E, \boldsymbol{X})$ or $(M, \boldsymbol{X})$ observed in all individual level datasets. Our approach of discretizing continuous variables can produce biased results and loss of efficiency. One solution is to build upon 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 Materials

Supplementary Materials contain web Appendices, Tables and Figures referenced in Sections 2, 3.2 and 4.

Acknowledgements

This study used the computational resources of the NIH HPC Biowulf cluster. We thank Clara Bodelon for access to the data used as the example.

References

Allendoerfer, C. B. (1974). *Calculus of several variables and differentiable manifolds.* Macmillan.

Bodelon, C., H. Oh, A. Derkach, J. N. Sampson, B. Sprague, P. Vacek, D. Weaver, S. Fan, M. Palakal, D. Papathomas, 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), 1–6.

Breslow, N. and R. Holubkov (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*(2), 447–461.

Buchanan, A. L., M. G. Hudgens, S. R. Cole, K. R. Mollan, P. E. Sax, E. S. Daar, A. A. Adimora, J. J. Eron, and M. J. Mugavero (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(4), 1193–1209.

Cao, W., A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734.

REFERENCES

Carroll, R., S. Wang, and C. Wang (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association 90*(429), 157–169.

Chatterjee, n., Y. Chen, P. Maas, and R. Carroll (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association 111*(513), 107–117.

Chen, Y. and H. Chen (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(3), 449–460.

Cheng, J., N. F. Cheng, Z. Guo, S. Gregorich, A. I. Ismail, and S. A. Gansky (2018). Mediation analysis for count and zero-inflated count data. *Statistical methods in medical research 27*(9), 2756–2774.

Cheng, W., J. G. Taylor, T. Gu, S. Tomlins, and B. Mukherjee (2019). Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 68*(1), 121–139.

Daniels, M. J., J. Roy, C. Kim, J. Hogan, and M. Perri (2012). Bayesian inference for the causal effect of mediation. *Biometrics 68*(4), 1028–1036.

Derkach, A., R. Pfeiffer, T. Chen, and J. Sampson (2019). High dimensional mediation analysis with latent variables. *Biometrics 75*(3), 745–756.

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

REFERENCES

Figueroa, J. D., R. M. Pfeiffer, D. A. Patel, L. Linville, L. A. Brinton, G. L. Gierach, X. R. Yang, D. Papathomas, D. Visscher, C. Mies, 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*(10).

Huang, Y.-T. (2019). Variance component tests of multivariate mediation effects under composite null hypotheses. *Biometrics 75*(4), 1191–1204.

Huang, Y.-T. and T. Cai (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics 72*(2), 563–574.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imai, K., L. Keele, and T. Yamamoto (2011). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Kundu, P., R. Tang, and N. Chatterjee (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika 106*(3), 567–585.

Lin, D. and Z. Tang (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics 89*(3), 354 – 367.

Lin, D. Y. and D. Zeng (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association 101*(473), 89–104.

Mavaddat, N., K. Michailidou, J. Dennis, M. Lush, L. Fachal, A. Lee, J. Tyrer, T. Chen,

Q. Wang, M. Bolla, et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics 104*(1), 21–34.

VanderWeele, T. J. (2014). A unification of mediation and interaction: a four-way decomposition. *Epidemiology (Cambridge, Mass.) 25*(5), 749.

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

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*(531), 1540–1554.

Yuan, K. and R. Jennrich (2000). Estimating equations with nuisance parameters: Theory and applications. *Annals of the Institute of Statistical Mathematics 52*(2), 343–350.

Zeng, D. and D. Lin (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association 109*(505), 371–383.

Zeng, S., S. Rosenbaum, S. C. Alberts, E. A. Archie, and F. Li (2021). Causal mediation analysis for sparse and irregular longitudinal data. *The Annals of Applied Statistics 15*(2), 747–767.

Zhang, H., L. Deng, M. Schiffman, J. Qin, and K. Yu (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika 107*(3), 689–703.

REFERENCES

[1]Department of Epidemiology and Biostatistics, MSKCC, New York, NY 10017, USA

E-mail: derkacha@mskcc.org

[2] Biostatistics Branch,DCEG, NCI, NIH, Rockville, MD 20852, USA

E-mail: joshua.sampson@nih.gov and pfeiffer@mail.nih.gov