Research Article

Myoung-jae Lee*

Direct, indirect, and interaction effects based on principal stratification with a binary mediator

https://doi.org/10.1515/jci-2023-0025
received April 27, 2023; accepted April 15, 2024

Abstract: Given a binary treatment and a binary mediator, mediation analysis decomposes the total effect of the treatment on an outcome variable into various sub-effects, and there appeared two-, three-, and four-way decompositions in the literature. Using "principal stratification" based on the potential mediator types, we consider sub-treatment effects for "mediative never-takers, compliers, defiers, and always takers." In this approach, although it is difficult to pick any one decomposition over the others in general, a particular three-way decomposition becomes well suited, which is thus advocated to use. We present identification conditions for the effects using conditional means, which is then followed by simple estimators that are applicable to any outcome variable (binary, count, continuous, etc.). We also provide simulation and empirical studies.

Keywords: direct effect, indirect effect, interaction effect, mediation, principal stratification

MSC 2020: 62D20

1 Introduction

Given a binary treatment D, a binary mediator M, and an outcome/response variable Y, the causal chain of interest in mediation analysis is

$$D \to Y \quad \text{and} \quad D \to M \to Y.$$

The total effect of D on Y consists of the direct effect of D on Y and the indirect effect through M. This is an important issue in many disciplines of science, as reviewed by MacKinnon et al. [1], Pearl [2], Imai et al. [3], TenHave and Joffe [4], Preacher [5], VanderWeele [6,7], and Nguyen et al. [8], among others.

The total effect of D on Y can be found in various ways, such as matching, regression adjustment and inverse probability weighting (see, e.g., Lee and Lee [9] and Choi and Lee [10] for reviews on treatment effect estimators). Traditionally, the total effect has been decomposed into sub-effects with linear structural-form (SF) models for M as a function of D and Y as a function of (D, M), e.g., with some α and β parameters,

$$M = \alpha_1 + \alpha_d D + \varepsilon \quad \text{and} \quad Y = \beta_1 + \beta_d D + \beta_m M + \beta_{dm} DM + U, \tag{1.1}$$

where (ε, U) are the error terms with $E(\varepsilon|D) = 0$ and $E(U|D, M) = 0$, α_d is the effect of D on M, β_d is the direct effect of D on Y, β_m is the direct effect of M on Y, and β_{dm} is the interaction effect of DM on Y. Here, $\beta_m \alpha_d$ is the "indirect effect" of D on Y through M, which is the "effect of M on Y" times the "effect of D on M." The standard

* **Corresponding author: Myoung-jae Lee,** Department of Finance, Accounting & Economics, University of Nottingham Ningbo China, Ningbo 315100, China; Department of Economics, Korea University, Seoul 02841, South Korea, e-mail: mjlee7867@gmail.com, myoungjae@korea.ac.kr

Baron-Kenny approach (Baron and Kenny [11]) does not include the interaction part $\beta_{dm}DM$, which turns out to be a major complicating term for total effect decomposition.

Linear SFs such as (1.1) are intuitive and easy to understand, but they are subject to specification errors and not grounded in the modern counterfactual causal framework. Once we adopt nonparametric approaches to avoid SF misspecifications and use potential versions of (M, Y), total effect decomposition is no longer straightforward as with (1.1). Note that when exogenous covariates X need to be controlled, an easy way to do it generalizing (1.1) is

$$M = \alpha_1(X) + \alpha_d(X)D + \varepsilon, \quad Y = \beta_1(X) + \beta_d(X)D + \beta_m(X)M + \beta_{dm}(X)DM + U, \tag{1.2}$$

where $\alpha(X)$s and $\beta(X)$s are the functions of X. Specifying them as linear, ordinary least-squares estimator (OLS) can be applied to (1.2). This approach turns out to be the main estimator proposed in this article, although (1.2) will be derived formally using the potential versions of M and Y.

Consider two potential versions M^d, $d = 0, 1$, of M corresponding to $D = 0, 1$, and the four potential responses Y^{dm} of Y for $D = 0, 1$ and $M = 0, 1$. Also, define the potential responses Y_d, $d = 0, 1$, corresponding to $D = 0, 1$ "when M is allowed to take its natural course given $D = d$":

$$Y_d \equiv Y^{d,M^d}.$$

Then, the mean/average total effect is

$$\text{Total effect} : E(Y_1 - Y_0) = E(Y^{1,M^1} - Y^{0,M^0}). \tag{1.3}$$

In the literature, two-way decompositions of the total effect appeared first, as can be seen in Pearl [12] and Robins [13], among others:

$$\begin{aligned} (a) &: E(Y^{1,M^1} - Y^{1,M^0}) + E\{Y^{1,M^0} - Y^{0,M^0}\}, \\ (b) &: E\{Y^{1,M^1} - Y^{0,M^1}\} + E(Y^{0,M^1} - Y^{0,M^0}). \end{aligned} \tag{1.4}$$

The terms in $\{\cdot\}$ are called the "natural direct effects" of D on Y, and the terms in $(\cdot)$ are the "natural indirect effects." In contrast, the "controlled direct effect" with $M = m$ is $E(Y^{1,m} - Y^{0,m})$, but there is no definition of "controlled indirect effect" that is generally agreed on. One glaring problem with (1.4) is that the decomposition is not unique: which one to take between (a) and (b)? Another problem is that the presence of the interaction effect is not clear in (1.4).

VanderWeele [14] proposed a three-way decomposition separating the aforementioned effects, and VanderWeele [15] proposed a four-way decomposition, which includes the existing two- and three-way decompositions in the literature as special cases by merging different terms in the four-way decomposition.

In this article, we define four subject types with (M^0, M^1):

$$\begin{aligned} \text{Mediative Never Takers (NT)} &: M^0 = 0, \quad M^1 = 0, \\ \text{Mediative ComPliers (CP)} &: M^0 = 0, \quad M^1 = 1, \\ \text{Mediative DeFiers (DF)} &: M^0 = 1, \quad M^1 = 0, \\ \text{Mediative Always Takers (AT)} &: M^0 = 1, \quad M^1 = 1. \end{aligned} \tag{1.5}$$

When D is endogenous with a binary instrumental variable δ for D, Imbens and Angrist [16] and Angrist et al. [17] classified the subjects using the potential treatments (D^0, D^1) corresponding to $\delta = 0, 1$, and the classification is analogous to (1.5), e.g., compliers are those with $(D^0 = 0, D^1 = 1)$. Frangakis and Rubin [18] called the classification based on (D^0, D^1) "principal stratification," and doing analogously, we call (1.5) "mediative principal stratification" based on (M^0, M^1). We address only binary M in this article; allowing non-binary M is not straightforward and thus left for a future research.

With many decompositions of the total effect in the literature, the question is which one to use. There seems no single best answer to this question, but once the mediator types are taken into account, the following three-way decomposition seems well suited:

$$E(Y^{10} - Y^{00}) + E\{(Y^{01} - Y^{00})(M^1 - M^0)\} + E\{(Y^{11} - Y^{01} - Y^{10} + Y^{00})M^1\}. \tag{M3M}$$

"M3M" stands for "Main 3-way Mediator-based decomposition." M3M appeared in VanderWeele [15] with different notation, who also referred to VanderWeele and Tchetgen Tchetgen [19]. The following explains the three terms/effects in M3M.

First, the *direct effect* $Y^{10} - Y^{00}$ occurs when D changes while the presence of M is nullified by $m = 0$ in Y^{dm}. Second, the *indirect effect through* M is the product of the "direct effect $Y^{01} - Y^{00}$ of M on Y (with D nullified by $d = 0$ in Y^{dm})" times the "effect $M^1 - M^0$ of D on M," which occurs only to CP and DF (with the opposite signs) as only they have $M^1 \neq M^0$. Third, the *interaction effect of DM* is the "net effect of DM," which is the "gross effect of DM" minus the "direct/partial" effects of D and M (Choi and Lee [20]):

$$\begin{aligned}\Delta Y^{\pm} &\equiv Y^{11} - Y^{01} - Y^{10} + Y^{00} = Y^{11} - Y^{00} - (Y^{01} - Y^{00}) - (Y^{10} - Y^{00}) \\ &= (\text{gross effect of } DM) - (\text{direct effect of } M) - (\text{direct effect of } D).\end{aligned} \tag{1.6}$$

This effect occurs only to CP and AT because only they have $DM = DM^1 = 1$, which is why M^1 (=1 only for CP and AT) is attached to $\Delta Y^{\pm}$ in M3M. Note that the interaction effect can be viewed either as the direct effect of D moderated by the level of M or as the direct effect of M moderated by the level of D.

Our approach based on "mediative principal stratification" looks at which effects are associated with which type, so that each type's contributions to the total effect suggest the appropriate decomposition of the total effect, as was seen just above. Our approach will further show that, despite the apparent "symmetry" in (1.4)(a) and (b), (1.4)(b) is better than (1.4)(a) because (1.4)(a) does not identify any effect in M3M, whereas (1.4)(b) identifies the indirect effect and the sum of the direct and interaction effects.

In the remainder of this article, Section 2 shows the details of M3M. Section 3 addresses identifying all sub-effects in M3M, and maps out our estimation strategy. Section 4 presents the effect estimators. Sections 5 and 6 provide simulation and empirical studies. Finally, Section 6 concludes this article. Appendix contains most proofs.

2 Main three-way mediator-based decomposition

In this section, first, we formally put forth M3M as a preferred decomposition and explain why. Second, we compare M3M to other decompositions in the literature. Third, we illustrate various decompositions, using simple SFs as in (1.1).

2.1 Mediator types and decompositions

Our starting point is the definitional equations for (Y_0, Y_1) in (1.3):

$$\begin{aligned}(Y_0\equiv)\ Y^{0,M^0} &= Y^{00} + (Y^{01} - Y^{00})M^0, \quad (Y_1\equiv)\ Y^{1,M^1} = Y^{10} + (Y^{11} - Y^{10})M^1, \\ Y^{0,M^1} &= Y^{00} + (Y^{01} - Y^{00})M^1, \quad Y^{1,M^0} = Y^{10} + (Y^{11} - Y^{10})M^0;\end{aligned} \tag{2.1}$$

these equalities can be verified by setting $M^0 = 0, 1$ and $M^1 = 0, 1$. We substitute these into the total effect (1.3) to obtain M3M, and we then explain why M3M is advocated.

Proposition 1. *The total effect* (1.3) *can be rewritten as our preferred three-way mediator-based decomposition M3M using the mediator types in* (1.5), *where* (i) *the interaction effect is for* $M^1 = 1$ *(CP or AT), (ii) the indirect effect is for* $M^1 - M^0 = 1$ *or* -1 *(CP or DF), and (iii) the direct effect is for every subject.*

Our approach using mediator types considers which effects are associated with which mediator types. For this, we examine the last term (interaction effect) in M3M first, followed by the indirect effect (the second term) and the direct effect (the first term).

First, the interaction effect (i.e., the effect of DM) requires $DM = DM^1 = 1$, which occurs only to the mediator types AT and CP because only they have $M^1 = 1$. Hence, M^1 in $\Delta Y^{\pm}M^1$ works as a "qualification indicator" for the interaction effect. Although we assume that M is binary to take advantage of the mediator types in (1.5), if M were non-binary, M^1 would work also as the level of the "causal intensity" of D. Interpreting $\Delta Y^{\pm}$ in (1.6) as the net effect of DM (i.e., interaction effect) follows from the definition that whatever is left in $Y^{11} - Y^{00}$ after the direct/partial effects of D and M are subtracted is the interaction effect.

Second, the indirect effect requires $M^0 \neq M^1$, which occurs only to CP and DF because only they have $M^0 \neq M^1$. Hence, $M^1 - M^0$ works as a qualification indicator for the indirect effect, much as M^1 does in the interaction effect $\Delta Y^{\pm}M^1$. $M^1 - M^0$ works also as the effect magnitude of D on M, so that when it is multiplied to the effect $Y^{01} - Y^{00}$ of M on Y, the indirect effect is obtained. For example, if D is education for health knowledge, Y is health, and M is exercising, then the CP's are those who exercise due to $D = 1$ with the indirect effect $(Y^{01} - Y^{00})(M^1 - M^0) = Y^{01} - Y^{00}$, whereas the DF's are those who stop exercising due to $D = 1$ with the indirect effect $Y^{00} - Y^{01}$ because of $M^1 - M^0 = -1$. Under $Y^{01} - Y^{00} > 0$, the indirect effect is positive for CP, but negative for DF because DF stops exercising; DF may not exist in this example though.

Third, since no qualification indicator involving M appears in the first term $Y^{10} - Y^{00}$ of M3M, the direct effect $Y^{10} - Y^{00}$ is associated with all types.

2.2 Comparisons to other decompositions

The most general decomposition yet is the four-way one in VanderWeele [15]:

$$\begin{aligned}
&\text{(i) interact no, mediate no} &&: \text{direct effect for all types} \quad E(Y^{10} - Y^{00}),\\
&\text{(ii) interact no, mediate yes} &&: \text{indirect effect for CP,DF} \quad E\{(Y^{01} - Y^{00})(M^1 - M^0)\},\\
&\text{(iii) interact yes, mediate no} &&: \text{interaction effect for AT,DF} \quad E(\Delta Y^{\pm}M^0),\\
&\text{(iv) interact yes, mediate yes} &&: \text{interaction effect for CP,DF} \quad E\{\Delta Y^{\pm}(M^1 - M^0)\},
\end{aligned} \tag{2.2}$$

using our notation. The type classifications are new, not in VanderWeele [15].

In (i), the direct effect of D needs neither interaction nor mediation, which is relevant for all types as no M appears there. In (ii) that is the indirect effect for CP and DF, the effect needs mediation ($M^1 - M^0$), but not interaction (no $\Delta Y^{\pm}$). In (iii) that is the interaction effect (with $\Delta Y^{\pm}$) for AT and DF because only they have $M^0 = 1$, no mediation is needed (i.e., no $M^1 - M^0$). In (iv) that is the interaction effect for CP and DF because only they have $M^1 \neq M^0$, the effect needs both interaction ($\Delta Y^{\pm}$) and mediation ($M^1 - M^0$). Note that VanderWeele [15] called (iii) "reference interaction" as M^0 appears, and (iv) "mediated interaction" as $M^1 - M^0$ appears. In the following, we compare (2.2) to other decompositions.

First, *the four-way decomposition reduces to M3M, when the two interaction effects (iii) and (iv) in* (2.2) *are merged* to yield the last term of M3M:

$$E(\Delta Y^{\pm}M^0) + E\{\Delta Y^{\pm}(M^1 - M^0)\} = E(\Delta Y^{\pm}M^1).$$

M3M is preferred to the four-way decomposition, because the interaction effect in M3M is only for CP and AT whereas (2.2)(iv) includes DF in the interaction effect despite that DF has $DM = DM^1 = 0$. Also, (2.2)(iv) can be taken as part of either the interaction effect (due to $\Delta Y^{\pm}$) or the indirect effect (due to $M^1 - M^0$), but M3M merges (2.2)(iv) into (iii) to obtain the interaction effect only for CP and AT, which is appropriate because only they have $DM = DM^1 = 1$.

Second, instead of merging (iii) and (iv) as was done just now, merge (ii) and (iv):

$$E\{(Y^{01} - Y^{00})(M^1 - M^0)\} + E\{(Y^{11} - Y^{10} - Y^{01} + Y^{00})(M^1 - M^0)\} = E\{(Y^{11} - Y^{10})(M^1 - M^0)\}.$$

This then yields a three-way decomposition similar to, yet different from, M3M:

$$E(Y^{10} - Y^{00}) + E\{(Y^{11} - Y^{10})(M^1 - M^0)\} + E(\Delta Y^{\pm}M^0). \tag{2.3}$$

The difference is that the middle indirect effect here has $Y^{11} - Y^{10}$, not $Y^{01} - Y^{00}$ in M3M, and the last interaction effect here is for AT and DF because $M^0 = 1$ only for AT and DF despite that the interaction effect should be zero for DF as they have $DM = DM^1 = 0$. Hence, (2.3) illustrates an inappropriate three-way decomposition.

Third, VanderWeele [14] proposed yet another three-way decomposition:

$$E(Y^{1,M^0} - Y^{0,M^0}) + E(Y^{0,M^1} - Y^{0,M^0}) + E\{\Delta Y^{\pm}(M^1 - M^0)\}, \tag{2.4}$$

using our notation; the three terms are direct, indirect, and interaction effects. In view of mediator types, however, the third term is inappropriate, because it is zero for AT due to $M^1 - M^0 = 0$ despite that the interaction effect do occur to AT in (2.2)(iii).

Fourth, substitute (2.1) into (1.4)(a):

$$\begin{aligned} &E[\{Y^{10} + (Y^{11} - Y^{10})M^1\} - \{Y^{10} + (Y^{11} - Y^{10})M^0\} + \{Y^{10} + (Y^{11} - Y^{10})M^0\} - \{Y^{00} + (Y^{01} - Y^{00})M^0\}] \\ &\quad = E\{(Y^{11} - Y^{10})(M^1 - M^0)\} + E(Y^{10} - Y^{00} + \Delta Y^{\pm}M^0). \end{aligned} \tag{2.5}$$

Hence, (1.4)(a) consists of two terms: the first is an indirect effect differing from (2.2)(ii) because Y^{11}–Y^{10} appears instead of Y^{01}–Y^{00}, and the second is the sum of the direct effect in (2.2)(i) and the interaction effect for AT and DF in (2.2)(iii) with the interaction effect for CP in (2.2)(iv) omitted. Hence, (1.4)(a) is inappropriate.

Now, substitute (2.1) into (1.4)(b):

$$\begin{aligned} &E[\{Y^{10} + (Y^{11} - Y^{10})M^1\} - \{Y^{00} + (Y^{01} - Y^{00})M^1\} + \{Y^{00} + (Y^{01} - Y^{00})M^1\} - \{Y^{00} + (Y^{01} - Y^{00})M^0\}] \\ &\quad = E(Y^{10} - Y^{00} + \Delta Y^{\pm}M^1) + E\{(Y^{01} - Y^{00})(M^1 - M^0)\}. \end{aligned} \tag{2.6}$$

This is the same as M3M, except that the direct and interaction effects of M3M are merged into the first term. Hence, despite the apparent "symmetry" between (1.4)(a) and (1.4)(b), (1.4)(b) is the better decomposition than (1.4)(a), as long as we are aware that the direct effect in (1.4)(b) is inclusive of the interaction effect.

If the control and treatment are switched, then (1.4)(b) can be written as

$$E\{Y^{0,M^0} - Y^{1,M^0}\} + E(Y^{1,M^0} - Y^{1,M^1}) = -(1.4)(a),$$

which raises the question whether (1.4)(b) becomes inappropriate with the switch. The answer is that our approach is not "symmetric" due to the definitions of CP and DF: the interaction effect becomes relevant now for DF and AT, as they have $M^0 = 1$ (i.e., $M = 1$ when actively treated). This makes (2.3), not M3M, the appropriate three-way decomposition, and then, the last display becomes the appropriate two-way decomposition (with the minus sign due to the treatment reversal).

2.3 Illustration of decompositions

To illustrate various decompositions, we use the following potential variable models:

$$\begin{aligned} M^d &= 1[0 < \alpha_1 + \alpha_d d + X'\alpha_x - e], \\ Y^{dm} &= \beta_1 + \beta_d d + \beta_m m + \beta_{dm} dm + X'\beta_x + U, \end{aligned} \tag{2.7}$$

where $1[A] \equiv 1$ if A holds and 0 otherwise, and (e, U) are the error terms independent of X, and $e \sim \text{Uni}(0,1)$ with $\text{Uni}(0,1)$ standing for the uniform distribution on $(0,1)$. The models are tightly specified, but they will help understand the aforementioned decompositions. This non-essential subsection may be skipped.

Assuming $0 < \alpha_1 + \alpha_d d + X'\alpha_x < 1$ for all X, it holds that

$$\begin{aligned} &E(M^d|X) = P(e < \alpha_1 + \alpha_d d + X'\alpha_x|X) = \alpha_1 + \alpha_d d + X'\alpha_x \\ &\Rightarrow M^d = \alpha_1 + \alpha_d d + X'\alpha_x + \varepsilon^d \\ &\text{where} \quad \varepsilon^d \equiv M^d - \alpha_1 - \alpha_d d - X'\alpha_x \quad \{E(\varepsilon^d|X) = 0 \text{ by construction}\}. \end{aligned} \tag{2.8}$$

The linear M^d model in (2.8) is a "reduced form (RF)" as opposed to its SF in (2.7). We use the linear RF for M^d in (2.8) and the linear SF for Y^{dm} in (2.7) in the following.

The Y^{dm} model gives $Y^{01} - Y^{00} = \beta_m$ and $Y^{11} - Y^{10} = \beta_m + \beta_{dm}$, which yield

$$E(Y^{0,M^0}) = E\{Y^{00} + (Y^{01} - Y^{00})M^0\} = \beta_1 + E(X')\beta_x + \beta_m\{\alpha_1 + E(X')\alpha_x\},$$
$$E(Y^{1,M^1}) = \beta_1 + \beta_d + E(X')\beta_x + (\beta_m + \beta_{dm})\{\alpha_1 + \alpha_d + E(X')\alpha_x\},$$

where (2.1) is used. Subtracting the former from the latter renders the total effect:

$$E(Y^{1,M^1} - Y^{0,M^0}) = \beta_d + \beta_m\alpha_d + \beta_{dm}\{\alpha_1 + \alpha_d + E(X')\alpha_x\}. \tag{2.9}$$

Different decompositions shuffle (2.9) in different ways as follows; note that

$$\Delta Y^{\pm} = Y^{11} - Y^{10} - (Y^{01} - Y^{00}) = \beta_m + \beta_{dm} - \beta_m = \beta_{dm}.$$

First, the four-way decomposition that is the sum of the four terms in (2.2) is

$$E(Y^{10} - Y^{00}) + E\{(Y^{01} - Y^{00})(M^1 - M^0)\} + E(\Delta Y^{\pm}M^0) + E\{\Delta Y^{\pm}(M^1 - M^0)\}$$
$$= \beta_d + \beta_m\alpha_d + \beta_{dm}\{\alpha_1 + E(X')\alpha_x\} + \beta_{dm}\alpha_d.$$

Here, the last term of (2.9) is split into $\beta_{dm}\{\alpha_1 + E(X')\alpha_x\}$ and $\beta_{dm}\alpha_d$.

Second, the three-way decomposition in M3M is the same as (2.9):

$$E(Y^{10} - Y^{00}) + E\{(Y^{01} - Y^{00})(M^1 - M^0)\} + E\{(Y^{11} - Y^{10} - Y^{01} + Y^{00})M^1\}$$
$$= \beta_d + \beta_m\alpha_d + \beta_{dm}E\{\alpha_1 + \alpha_d + E(X')\alpha_x\},$$

which are the direct, indirect, and interaction effects, respectively.

Third, for (1.4)(a), substitute (2.7) and (2.8) into (2.5):

$$E\{(Y^{11} - Y^{10})(M^1 - M^0)\} + E(Y^{10} - Y^{00} + \Delta Y^{\pm}M^0) = (\beta_m + \beta_{dm})\alpha_d + [\beta_d + \beta_{dm}\{\alpha_1 + E(X')\alpha_x\}].$$

This is the same as the total effect in (2.9), but both terms here include the interaction effect β_{dm}, revealing again why the two-way decomposition (1.4)(a) is inappropriate.

Fourth, for (1.4)(b), substitute (2.7) and (2.8) into (2.6):

$$E(Y^{10} - Y^{00} + \Delta Y^{\pm}M^1) + E\{(Y^{01} - Y^{00})(M^1 - M^0)\} = [\beta_d + \beta_{dm}\{\alpha_1 + \alpha_d + E(X')\alpha_x\}] + \beta_m\alpha_d.$$

The second term is the indirect effect of M3M, whereas the first term is the sum of the direct and interaction effects, revealing again why (1.4)(b) is better than (1.4)(a).

3 Identification and estimation strategy

Our identification conditions for M3M with X are ("$\perp\!\!\!\perp$" stands for independence):

C(a) : (i) $D \perp\!\!\!\perp (M^0, M^1)|X$ (ii) $D \perp\!\!\!\perp (Y^{00}, Y^{01}, Y^{10}, Y^{11})|X$;
C(b) : $(M^0, M^1) \perp\!\!\!\perp (Y^{00}, Y^{01}, Y^{10}, Y^{11})|(D, X)$;
C(c) : $0 < P(D = d, M = m|X)$ for all $d, m = 0, 1$ and X.

Conditions C(a) and C(b) are the "ignorability" of confounders in the treatment-mediator, treatment-outcome, and mediator-outcome relationships. C(a) and C(b) operate in two stages: as D precedes M, which, in turn, precedes Y, the first stage is D being independent of all potential future variables given X, and the second stage is (M^0, M^1) being independent of all potential future versions of Y given (D, X). C(c) is a support-overlap condition for $D|X$ and $M|(D, X)$; for example, C(c) implies

$$P(D = d|X) = P(D = d, M = 0|X) + P(D = d, M = 1|X) > 0, \quad \text{for } d = 0, 1,$$
$$P(M = m|D = d, X) = P(D = d, M = m|X)/P(D = d|X) > 0, \quad \text{for } d, m = 0, 1.$$

Slightly different conditions from C(a) and C(b) appeared in Imai et al. [3]:

$$D \perp\!\!\!\perp (M^d, Y^{d'm})|X \quad \text{and} \quad M^d \perp\!\!\!\perp Y^{d'm}|(D, X), \quad \text{for all } d, d', m = 0, 1$$

where the joint distributions of (M^0, M^1) and $(Y^{00}, Y^{01}, Y^{10}, Y^{11})$ do not appear, differently from C(a) and C(b). Also, Petersen et al. [21] assumed

$$D \perp\!\!\!\perp M^d|X, \quad D \perp\!\!\!\perp Y^{dm}|X, \quad M \perp\!\!\!\perp Y^{dm}|(D, X), \quad \text{for all } d, m = 0, 1.$$

Using marginal independence instead of joint independence, we can relax C(a) and C(b), but we continue to assume C(a) and C(b) for simplicity. In the following, we present "causal reduced forms (CRF's)" for M and Y, which form the basis for identification.

Proposition 2. *Under C(a) to C(c), the following CRF's hold for M and Y*:

$$\begin{aligned} &M = \psi_1(X) + \psi_d(X)D + U_m, \quad \textit{where} \\ &U_m \equiv M - E(M|D, X), \quad \psi_1(X) \equiv E(M^0|X), \quad \psi_d(X) \equiv E(M^1 - M^0|X); \end{aligned} \tag{3.1}$$

$$\begin{aligned} &Y = \mu_1(X) + \mu_d(X)D + \mu_m(X)M + \mu_{dm}(X)DM + U_y, \\ &U_y \equiv Y - E(Y|D, M, X), \quad \mu_1(X) \equiv E(Y^{00}|X), \quad \mu_d(X) \equiv E(Y^{10} - Y^{00}|X), \\ &\mu_m(X) \equiv E(Y^{01} - Y^{00}|X), \quad \textit{and} \quad \mu_{dm}(X) \equiv E(\Delta Y^{\pm}|X), \end{aligned} \tag{3.2}$$

where $\psi_d(X)$, $\mu_d(X)$, $\mu_m(X)$, and $\mu_{dm}(X)$ are the X-conditional causal effects of interest.

Regarding (3.1), it holds for any form of M as long as $M^1 - M^0$ makes sense. However, since M3M is based on binary M, we continue to assume binary M. There are two "cells" $D = 0, 1$, and $\psi_1(X)$ and $\psi_d(X)$ are nonparametrically identified using

$$E(M|X, D = 0) = E(M^0|X) = \psi_1(X), \text{ and } E(M|X, D = 1) = E(M^1|X) = \psi_1(X) + \psi_d(X).$$

This point can be understood by considering a nonparametric estimation of M on X for the $D = 0, 1$ groups, separately.

Analogously to (3.1), (3.2) holds also for any form of Y, as long as $Y^{10} - Y^{00}$, $Y^{01} - Y^{00}$, and $\Delta Y^{\pm}$ make sense; For example, Y can be binary, continuous, etc. There are four cells formed by $D = 0, 1$ and $M = 0, 1$, and analogously to the identification of $\{\psi_1(X), \psi_d(X)\}$, $\{\mu_1(X), \mu_d(X), \mu_m(X), \mu_{dm}(X)\}$ are nonparametrically identified using the X-conditional means on the cells.

A model is a SF, if it is a data-generating process with parameters of interest governing the behaviors of subjects. A model is a RF, if it is derived from some SF's; e.g., a SF for Y has D and M on the right-hand side, and substituting out the SF for M yields the RF for Y with only D on the right-hand side. The parameters in a RF are not of direct interest, as they are functions of the underlying SF parameters. CRF falls between SF and RF, as it is a derived form or RF but with causal parameters of interest, such as $\mu_d(X)$, $\mu_m(X)$, and $\mu_{dm}(X)$ in (3.2). CRF's similar to (3.1) and (3.2) appeared in Lee [22,23], Mao and Li [24], Choi et al. [25], and Lee et al. [26].

To understand the effect decomposition better, substitute (3.1) into (3.2):

$$Y = \mu_1(X) + \mu_m(X)\psi_1(X) + [\mu_d(X) + \mu_m(X)\psi_d(X) + \mu_{dm}(X)\{\psi_1(X) + \psi_d(X)\}]D + \{\mu_m(X) + \mu_{dm}(X)D\}U_m + U_y. \tag{3.3}$$

The slope of D is the "X-conditioned total effect" consisting of direct ($\mu_d(X)$), indirect ($\mu_m(X)\psi_d(X)$) and interaction ($\mu_{dm}(X)\{\psi_1(X) + \psi_d(X)\}$) effects. Compare these to the constant-effect versions in (2.9): $\beta_d + \beta_m\alpha_d + \beta_{dm}\{\alpha_1 + E(X')\alpha_x + \alpha_d\}$.

If desired, $\{\psi_1(X), \psi_d(X), \mu_1(X), \mu_d(X), \mu_m(X), \mu_{dm}(X)\}$ in (3.1) and (3.2) can be estimated nonparametrically. However, more practical would be specifying those functions as, e.g., linear functions of X, and then, apply OLS to the resulting M and Y models. The details of this approach will be seen in the next section.

The OLS just mentioned is reminiscent of the traditional approach with (1.1), which raises the question: how much is the aforementioned OLS different from the traditional OLS to (1.1)? The answer is that there are three critical differences.

First, whereas the SF's in (1.1), i.e., the data-generating processes, hold only for continuous M and Y in general, the aforementioned CRF's for M and Y hold more generally for any forms of M and Y.

Second, (1.1) can be generalized to account for effect heterogeneity as in (1.2). However, proceeding with the SF's in (1.2) makes it unclear what kind of direct, indirect, and interaction effects are actually estimated: e.g., are they (2.3) or M3M?

Third, if one starts off with (1.2), then it is not clear whether $\alpha_d(X)$, $\beta_d(X)$, $\beta_m(X)$, and $\beta_{dm}(X)$ in (1.2) should be restricted or not. In contrast, the definitions of the unknown functions of X in (3.1) and (3.2) show how they might be restricted. For example, since $\psi_1(X) \equiv E(M^0|X)$, for binary M as in our setup, $\psi_1(X)$ cannot be specified just as $X'\alpha$ for a parameter α; rather, $\psi_1(X) = \Phi(X'\alpha)$ is more suitable for a distribution function $\Phi(\cdot)$. Another example is $\mu_d(X) \equiv E(Y^{10} - Y^{00}|X)$: if Y is binary, $E(Y^{10} - Y^{00}|X)$ should be bounded by $[-1, 1]$, which can be accommodated by a smooth function with the range on $[-1, 1]$, such as the arctan function or $\Phi(X'\alpha_{10}) - \Phi(X'\alpha_{00})$, where $\Phi(X'\alpha_{10})$ is for $E(Y^{10}|X)$ and $\Phi(X'\alpha_{00})$ is for $E(Y^{00}|X)$. If these nonlinear functions are used, then the aforementioned OLS should be replaced by a nonlinear least-squares estimator, which will be further discussed in the next section.

4 Effect estimators using OLS and generalized method of moment (GMM)

To estimate the effects in M3M, we linearly approximate all $\psi(X)$ and $\mu(X)$ terms in the CRF's (3.1) and (3.2), and then apply OLS to (3.1) and (3.2). This is summarized in Proposition 3, which is followed by discussions on nonlinear estimation when some $\psi(X)$ and $\mu(X)$ terms are nonlinearly specified.

Let $X_\alpha, X_1, X_d, X_m, X_{dm}$ consist of elements of X and their functions, with X_j of dimension $k_j \times 1$, $j = \alpha, 1, d, m, dm$. Linearly approximate the functions of X in (3.1):

$$M = \alpha_1'X_\alpha + \alpha_d'X_\alpha D + U_m = \alpha_m'Q_m + U_m, \quad \alpha_m \equiv (\alpha_1', \alpha_d')', \quad Q_m \equiv (X_\alpha', X_\alpha' D)'; \tag{4.1}$$

$\alpha_1'X_\alpha$ is for $\psi_1(X)$ in (3.1) and $\alpha_d'X_\alpha$ is for $\psi_d(X)$. For simplicity, we use the same X_α in $\alpha_1'X_\alpha$ and $\alpha_d'X_\alpha$, but we can certainly use different covariates, if desired.

Doing analogously for (3.2), we have

$$\begin{aligned} Y &= \beta_1'X_1 + \beta_d'X_d D + \beta_m'X_m M + \beta_{dm}'X_{dm} DM + U_y = \beta_y'Q_y + U_y, \\ \beta_y &\equiv (\beta_1', \beta_d', \beta_m', \beta_{dm}')', \quad Q_y = (X_1', X_d' D, X_m' M, X_{dm}' DM)', \end{aligned} \tag{4.2}$$

where $\beta_1'X_1$, $\beta_d'X_d$, $\beta_m'X_m$, and $\beta_{dm}'X_{dm}$ are for $\mu_1(X)$, $\mu_d(X)$, $\mu_m(X)$, and $\mu_{dm}(X)$.

One might use different covariates X_α, X_1, X_d, X_m, and X_{dm} as in (4.1) and (4.2), but this approach would be following the conventional SF view as in (1.2). Rather, since our CRF's are not SF's and the X-conditional causal parameters are of RF type (e.g., $\psi_d(X) \equiv E(M^1 - M^0|X)$ is a RF function), it is better to control for all X to ensure C(a) to C(c), i.e., setting $X = X_\alpha = X_1 = X_d = X_m = X_{dm}$ in (4.1) and (4.2) would be fine, as will be done in our simulation and empirical studies.

For asymptotic inference, we condition on $\bar{X}$; an upper bar denotes the sample average. This is to ignore errors of the form $\overline{X_m X_\alpha'} - E(X_m X_\alpha')$ relevant for the indirect and interaction effects, as accounting for such errors requires vectorizing the matrix $\overline{X_m X_\alpha'} - E(X_m X_\alpha')$ — unnecessary complications. What is gained by conditioning on $\bar{X}$ is ease in doing asymptotic inference, and what is lost is some "external validity," as the findings conditioned on $\bar{X}$ apply only to X-fixed designs in principle. Our simulation study with random X will demonstrate, however, that not accounting for errors of the form $\overline{X_m X_\alpha'} - E(X_m X_\alpha')$ makes little difference.

Let $0_{a\times b}$ be the $a \times b$ null vector and N be the sample size of independent and identically distributed observations; "$\hat{\beta}_d$" denotes an estimator for β_d. All three effects in M3M are found by the two OLS's of M on Q_m and Y on Q_y. Appendix provides missing details in Proposition 3. Recall (3.3): the indirect effect is $\mu_m(X) \cdot \psi_d(X)$ and the interaction effect is $\mu_{dm}(X) \cdot \{\psi_1(X) + \psi_d(X)\}$.

Proposition 3.

(*i*) *The direct effect estimator is* $\bar{X}_d'\hat{\beta}_d$ *with*

$$\sqrt{N}\bar{X}_d'(\hat{\beta}_d - \beta_d) \to^d N(0, \Lambda_d), \quad \hat{\Lambda}_d \equiv \frac{1}{N}\sum_i \hat{\lambda}_{di}^2 \to^p \Lambda_d,$$

$$\hat{\lambda}_{di} \equiv \hat{G}_d\left(\frac{1}{N}\sum_i Q_{yi}Q_{yi}'\right)^{-1} Q_{yi}\hat{U}_{yi}, \quad \hat{G}_d \equiv (0_{1\times k_1}, \bar{X}_d', 0_{1\times(k_m+k_{dm})}), \quad \hat{U}_{yi} \equiv Y_i - \hat{\beta}_y'Q_{yi}.$$

(*ii*) *The indirect effect estimator is* $\hat{\beta}_m'\overline{X_m X_\alpha'}\hat{\alpha}_d$ *with*

$$\sqrt{N}(\hat{\beta}_m'\overline{X_m X_\alpha'}\hat{\alpha}_d - \beta_m'\overline{X_m X_\alpha'}\alpha_d) \to^d N(0, \Lambda_m), \quad \hat{\Lambda}_m \equiv \frac{1}{N}\sum_i \hat{\lambda}_{mi}^2 \to^p \Lambda_m,$$

$$\hat{\lambda}_{mi} \equiv \hat{G}_{m1}\left(\frac{1}{N}\sum_i Q_{yi}Q_{yi}'\right)^{-1} Q_{yi}\hat{U}_{yi} + \hat{G}_{m2}\left(\frac{1}{N}\sum_i Q_{mi}Q_{mi}'\right)^{-1} Q_{mi}\hat{U}_{mi},$$

$$\hat{G}_{m1} \equiv (0_{1\times(k_1+k_d)}, \hat{\alpha}_d'\overline{X_\alpha X_m'}, 0_{1\times k_{dm}}), \quad \hat{G}_{m2} \equiv (0_{1\times k_\alpha}, \hat{\beta}_m'\overline{X_m X_\alpha'}), \quad \hat{U}_{mi} \equiv M_i - \hat{\alpha}_m'Q_{mi}.$$

(*iii*) *The interaction effect estimator is* $\hat{\beta}_{dm}'\overline{X_{dm}X_\alpha'}(\hat{\alpha}_1 + \hat{\alpha}_d)$ *with*

$$\sqrt{N}\{\hat{\beta}_{dm}'\overline{X_{dm}X_\alpha'}(\hat{\alpha}_1 + \hat{\alpha}_d) - \beta_{dm}'\overline{X_{dm}X_\alpha'}(\alpha_1 + \alpha_d)\} \to^d N(0, \Lambda_{dm}),$$

$$\hat{\Lambda}_{dm} \equiv \frac{1}{N}\sum_i \hat{\lambda}_{dmi}^2 \to^p \Lambda_{dm},$$

$$\hat{\lambda}_{dmi} \equiv \hat{G}_{dm1}\left(\frac{1}{N}\sum_i Q_{yi}Q_{yi}'\right)^{-1} Q_{yi}\hat{U}_{yi} + \hat{G}_{dm2}\left(\frac{1}{N}\sum_i Q_{mi}Q_{mi}'\right)^{-1} Q_{mi}\hat{U}_{mi},$$

$$\hat{G}_{dm1} \equiv (0_{1\times(k_1+k_d+k_m)}, (\hat{\alpha}_1 + \hat{\alpha}_d)'\overline{X_\alpha X_{dm}'}), \quad \hat{G}_{dm2} \equiv (\hat{\beta}_{dm}'\overline{X_{dm}X_\alpha'}, \hat{\beta}_{dm}'\overline{X_{dm}X_\alpha'}).$$

(iv) *The total effect estimator is* $\bar{X}_d'\hat{\beta}_d + \hat{\beta}_m'\overline{X_m X_\alpha'}\hat{\alpha}_d + \hat{\beta}_{dm}'\overline{X_{dm}X_\alpha'}(\hat{\alpha}_1 + \hat{\alpha}_d)$*, which is asymptotically normal with the variance estimated by* $N^{-1}\sum_i(\hat{\lambda}_{di} + \hat{\lambda}_{mi} + \hat{\lambda}_{dmi})^2$.

For continuous Y, linear approximations for $\{\mu_1(X), \mu_d(X), \mu_m(X), \mu_{dm}(X)\}$ should be fine, but linear approximations for $\psi_1(X)$ and $\psi_d(X)$ can result in biases because M is binary; the same is true of $\{\mu_1(X), \mu_d(X), \mu_m(X), \mu_{dm}(X)\}$ if Y is binary. In these cases, as was already noted, using functions with the ranges on [0,1] or $[-1, 1]$ would improve the approximations. However, the price to pay is the ensuing complications in estimation, because a nonlinear estimator such as GMM (Hansen [27]) is necessary.

For GMM to the M-CRF, consider a nonlinear moment condition:

$$E\{\theta(Z; \alpha)\} = 0, \quad \theta(Z; \alpha) \equiv [M - \Phi(\alpha_c'X_\alpha) - \{\Phi(\alpha_t'X_\alpha) - \Phi(\alpha_c'X_\alpha)\}D](X_\alpha', X_\alpha'D)', \tag{4.3}$$

where $Z \equiv (M, X_\alpha', D)'$, $\Phi(\alpha_c'X_\alpha) = E(M^0|X_\alpha)$, $\Phi(\alpha_t'X_\alpha) \equiv E(M^1|X_\alpha)$, and $\alpha \equiv (\alpha_c', \alpha_t')'$; the parameters α_c and α_t are for the control and treatment groups, when M is taken as the outcome. Appendix provides the GMM details.

Once the GMM for the M-CRF is done, $E(M^1 - M^0|X_\alpha) = X_\alpha'\alpha_d$ for the indirect effect and $E(M|X, D = 1) = E(M^1|X) = X_\alpha'(\alpha_1 + \alpha_d)$ for the interaction effect in Proposition 3 should be replaced by the estimates for $\Phi(\alpha_t'X_\alpha) - \Phi(\alpha_c'X_\alpha)$ and $\Phi(\alpha_t'X_\alpha)$, respectively. Appendix also addresses binary Y, for which $E(Y^{jk}|X) = \Phi(X'\beta_{jk})$ with a parameter β_{jk}, $j, k = 0, 1$, is adopted; if Y is a count or zero-censored, then $E(Y^{jk}|X) = \exp(X'\beta_{jk})$ can be used instead.

The asymptotic distributions for the effects with nonlinear approximations can be derived as in Proposition 3. However, since Y can take diverse forms, finding the asymptotic distributions of the effects for all forms of Y is cumbersome. Instead, we use bootstrap for asymptotic inference.

5 Simulation study

Recalling (2.7), we use four designs in our simulation study with D randomized, $P(D = 0) = P(D = 1) = 0.5$, $N = 250$, 1,000, and 5,000 simulation repetitions:

$$\begin{aligned}
&\text{Design 1:} \quad M^d = 1[0 < \alpha_1 + \alpha_d d + X'\alpha_x - \text{Uni}(0,1)], X \sim \text{Uni}(0,1), \text{continuous } Y^{dm};\\
&\text{Design 2:} \quad M^d = 1[0 < \alpha_1 + \alpha_d d + X'\alpha_x - \text{Uni}(0,1)], X \sim \text{Uni}(0,1), \text{probit } Y^{dm};\\
&\text{Design 3:} \quad M^d = 1[0 < \alpha_1 + \alpha_d d + X'\alpha_x + N(0,1)], X \sim N(0, 2^2), \text{continuous } Y^{dm};\\
&\text{Design 4:} \quad M^d = 1[0 < \alpha_1 + \alpha_d d + X'\alpha_x + N(0,1)], X \sim N(0, 2^2), \text{probit } Y^{dm}.
\end{aligned}$$

The error terms for M^d and Y^{dm} are independent of each other and X. "Probit Y^{dm}" means $Y^{dm} = 1[0 <$ continuous $Y^{dm}]$, where "continuous Y^{dm}" is the Y^{dm} in (2.7) with $U \sim N(0,1)$. In Designs 1 and 2, $E(M^d|X)$ is linear as was seen in (2.8).

As for the parameter values, we set

$$\alpha_1 = 0, \quad \alpha_d = \alpha_x = 0.5; \quad \beta_1 = 0, \quad \beta_d = \beta_m = \beta_{dm} = 0.5, \quad \text{and} \quad \beta_x = -1.$$

$\beta_x = -1$ prevents binary Y[11] from having too many 0s. We generate M and Y with

$$M = (1 - D)M^0 + DM^1, \quad Y = (1 - D)(1 - M)Y^{00} + (1 - D)MY^{01} + D(1 - M)Y^{10} + DMY^{11}.$$

In Design 1, the true effects are all constant: omitting "effect,"

$$\text{total} = \text{direct} + \text{indirect} + \text{interaction} = \beta_d + \beta_m\alpha_d + \beta_{dm}(\alpha_1 + \alpha_d + 0.5\alpha_x),$$

where 0.5 comes from $E(X) = E\{\text{Uni}(0,1)\} = 0.5$. In the other designs, however, the true effects are heterogeneous, and they are found numerically. For example, the true direct, indirect, and interaction effects in Design 2 are, recalling $\Delta Y^{\pm}$ for interaction effect,

$$\begin{aligned}
&E\{\Phi(\beta_1 + \beta_d + X'\beta_x) - \Phi(\beta_1 + X'\beta_x)\}, \quad E\{\Phi(\beta_1 + \beta_m + X'\beta_x) - \Phi(\beta_1 + X'\beta_x)\}\alpha_d,\\
&E[\{\Phi(\beta_1 + \beta_d + \beta_m + \beta_{dm} + X'\beta_x) - \Phi(\beta_1 + \beta_d + X'\beta_x) - \Phi(\beta_1 + \beta_m + X'\beta_x) + \Phi(\beta_1 + X'\beta_x)\}\cdot(\alpha_1 + \alpha_d + X'\alpha_x)].
\end{aligned}$$

The effects are complicated for designs 3 and 4 due to the $N(0,1)$ error in M^d.

We use two OLS's with $X = X_\alpha = X_1 = X_d = X_m = X_{dm}$: OLS_{v1} uses the random variable X as the single covariate, and OLS_{v2} uses X^2 additionally. Since OLS_{v2} uses one more covariate than OLS_{v1} does, OLS_{v2} is likely to be less biased but more dispersed than OLS_{v1}. Only in Design 4, we use "OLS_{v3}" that uses one more covariate $\Phi(X)$ than OLS_{v2} does to improve the linear approximation; GMM is also used in Design 4.

Table 1 presents the Design 1 (left-half) and Design 2 (right-half) results. Each entry has four numbers: |Bias|, standard deviation (Sd), root-mean-squared error (Rmse), and the average of 5,000 asymptotic Sd's to see

Table 1: |Bias/effect|, Sd/|effect|, (Rmse/|effect|), and asymptotic Sd/|effect|

	Design 1, $N = 250$	**Design 1, $N = 1{,}000$**	**Design 2, $N = 250$**	**Design 2, $N = 1{,}000$**
OLS_{v1}				
tot	0.00 0.12 (0.12) 0.12	0.00 0.06 (0.06) 0.06	0.00 0.15 (0.15) 0.14	0.00 0.07 (0.07) 0.07
dir	0.00 0.50 (0.50) 0.47	0.00 0.24 (0.24) 0.24	0.00 0.67 (0.67) 0.63	0.01 0.32 (0.32) 0.32
ind	0.02 0.53 (0.53) 0.50	0.01 0.25 (0.25) 0.25	0.01 0.70 (0.70) 0.65	0.00 0.33 (0.33) 0.33
int	0.01 0.73 (0.73) 0.70	0.00 0.35 (0.35) 0.35	0.00 1.14 (1.14) 1.08	0.00 0.54 (0.54) 0.53
OLS_{v2}				
tot	0.00 0.12 (0.12) 0.12	0.00 0.06 (0.06) 0.06	0.00 0.15 (0.15) 0.14	0.00 0.07 (0.07) 0.07
dir	0.00 0.64 (0.64) 0.51	0.00 0.27 (0.27) 0.26	0.02 0.82 (0.82) 0.68	0.00 0.35 (0.35) 0.34
ind	0.00 0.67 (0.67) 0.55	0.00 0.28 (0.28) 0.27	0.02 0.89 (0.89) 0.71	0.00 0.36 (0.36) 0.35
int	0.00 0.93 (0.93) 0.77	0.01 0.39 (0.39) 0.38	0.02 1.45 (1.45) 1.19	0.00 0.59 (0.59) 0.58
tru	1.125, 0.500, 0.250, 0.375		0.395, 0.184, 0.092, 0.118	

OLS_{v1} uses X & OLS_{v2} uses (X, X^2) for X-heterogeneous effect linear approximations; tot: total effect; dir: direct; ind: indirect; int: interaction; tru: true effect.

how accurate the variance formulas in Proposition 3 are, compared with the actual simulation Sd. Since the effects vary across the designs, we divide each number by the absolute effect magnitude for standardization.

In Design 1 with $N = 250$, all biases are almost zero, and OLS_{v1} does better than OLS_{v2} that is more dispersed than OLS_{v1}. With $N = 1{,}000$, both OLS's improve, and the performance differences narrow by much. In Design 2 with binary Y, OLS_{v1} still does better than OLS_{v2}. The second and fourth numbers in each entry of Table 1 are almost the same when $N = 1{,}000$, showing that the asymptotic variance formulas are accurate; for this, not accounting for the errors of the form $\bar{X} - E(X)$ hardly matters.

Table 2 presents the Design 3 (left-half) and Design 4 (right-half) results. In Design 3 with continuous Y, OLS_{v1} still does better than OLS_{v2} as in Table 1. Both OLS_{v1} and OLS_{v2} are little biased in Design 3, despite that only linear functions are used for the nonlinear $E(M^0|X)$ and $E(M^1|X)$.

In Design 4, we set $N = 500$, 2,000 because the GMM often incurred a singularity problem with $N = 250$. OLS_{v1} has much larger biases than OLS_{v2}. Since even the biases of OLS_{v2} are not negligible, to see if they can be reduced, we use $\Phi(X)$ as an extra covariate, which is OLS_{v3}. Indeed, the biases of OLS_{v3} are much smaller than those of OLS_{v1} and OLS_{v2}, and OLS_{v3} does better than OLS_{v1} and OLS_{v2} in terms of Rmse when $N = 2{,}000$. We also applied the GMM described in Appendix to the M and Y CRF's. GMM_{v1} using only X as OLS_{v1} does performs best, with its biases clearly decreasing when N quadruples to 2,000, whereas the OLS biases do not.

In summary, first, OLS_{v2} with two covariates (X, X^2) has higher Sd's than OLS_{v1} with only X, but depending on the true model, OLS_{v2} can be better than OLS_{v1} for a large N when the Sd of OLS_{v2} becomes much smaller while OLS_{v1} remains highly biased. Second, the asymptotic Sd formulas of our estimators work well, and are not affected by ignoring errors of the form $\bar{X} - E(X)$. Third, the binary nature of M seems to hardly matter, but the binary nature of Y does, for which GMM with nonlinear approximations is better than OLS with linear approximations. For this, of course, the nonlinear functions in GMM should be correctly specified.

Table 2: |Bias/effect|, Sd/effect, (Rmse/|effect|), and asymptotic Sd/|effect|

	Design 3, $N = 250$	**Design 3, $N = 1{,}000$**	**Design 4, $N = 500$**	**Design 4, $N = 2{,}000$**
OLS_{v1}				
tot	0.00 0.15 (0.15) 0.15	0.00 0.07 (0.07) 0.07	0.00 0.19 (0.19) 0.18	0.00 0.09 (0.09) 0.09
dir	0.01 0.53 (0.53) 0.50	0.00 0.25 (0.25) 0.25	0.74 0.88 (1.15) 0.84	0.72 0.42 (0.84) 0.42
ind	0.01 0.63 (0.63) 0.63	0.01 0.29 (0.29) 0.30	0.28 0.61 (0.67) 0.61	0.28 0.29 (0.41) 0.29
int	0.02 0.78 (0.78) 0.74	0.00 0.37 (0.37) 0.37	0.92 1.10 (1.43) 1.06	0.92 0.53 (1.06) 0.53
OLS_{v2}				
tot	0.00 0.16 (0.16) 0.15	0.00 0.07 (0.07) 0.07	0.00 0.18 (0.18) 0.18	0.00 0.09 (0.09) 0.09
dir	0.02 0.71 (0.71) 0.62	0.01 0.32 (0.32) 0.31	0.15 1.06 (1.07) 0.97	0.15 0.50 (0.52) 0.49
ind	0.01 0.70 (0.70) 0.71	0.01 0.30 (0.30) 0.31	0.14 0.90 (0.91) 0.87	0.08 0.43 (0.44) 0.42
int	0.03 1.09 (1.09) 0.94	0.01 0.48 (0.48) 0.46	0.23 1.38 (1.40) 1.26	0.22 0.65 (0.68) 0.64
OLS_{v3}				
tot			0.00 0.18 (0.18) 0.18	0.00 0.09 (0.09) 0.09
dir			0.04 0.98 (0.98) 0.82	0.05 0.36 (0.36) 0.35
ind			0.05 0.64 (0.64) 0.65	0.04 0.29 (0.29) 0.29
int			0.06 1.31 (1.31) 1.09	0.07 0.47 (0.48) 0.46
GMM_{v1}				
tot			0.01 0.18 (0.18) 0.00	0.01 0.09 (0.09) 0.00
dir			0.04 0.66 (0.66) 0.00	0.02 0.31 (0.31) 0.00
ind			0.00 0.57 (0.57) 0.00	0.00 0.27 (0.27) 0.00
int			0.06 0.82 (0.83) 0.00	0.03 0.39 (0.40) 0.00
tru	0.888, 0.500, 0.069, 0.319		0.168, 0.088, 0.015, 0.064	

OLS_{v1}, OLS_{v2} and OLS_{v3} use X, (X, X^2) and $\{X, X^2, \Phi(X)\}$, respectively; tot: total effect; dir: direct; ind: indirect; int: interaction; tru: true effect.

Table 3: Effect (t-value) of being black on wage with college education M

	OLS_{v0} (tv)	OLS_{v1}(tv)	OLS_{v2}(tv)	GMM_{v1} (tv)
Total effect	−0.319 (−17)	−0.223 (−9.0)	−0.220 (−8.5)	−0.222 (−9.0)
Direct	−0.336 (−15)	−0.242 (−7.7)	−0.248 (−8.0)	−0.242 (−7.7)
Indirect	−0.028 (−5.1)	−0.029 (−5.2)	−0.023 (−3.5)	−0.029 (−6.1)
Interaction	0.045 (3.6)	0.049 (3.2)	0.051 (3.1)	0.049 (3.4)

OLS_{v0} is the conventional constant-effect OLS with "intercept" $X'\beta_x$; OLS_{v1} approximates X-heterogeneous effects with linear functions of X; OLS_{v2} uses additionally the interactions between age & the other covariates; GMM_{v1} taking $M \in [0,1]$ into account uses the same set of X as in OLS_{v1}.

6 Empirical analysis

Our empirical analysis uses the National Longitudinal Survey data in Card [28], which have been used also in Tan [29] and Wang et al. [30], among others. With $N =$ 3,010, Y is ln(wage in 1976), D is the dummy for black, M is the dummy for some college education (i.e., schooling years being 13 or greater), and X consists of age, dummies ("r2, r3, ...") for 8 residence regions in 1966, dummy for living in a standard metropolitan statistical area (SMSA_{66}) in 1966, dummy for living in SMSA in 1976 ("SMSA"), and dummy for living in South in 1976 ("south"). In the original data, there were nine residence region dummies, but the dummy for region 8 was dropped due to a singularity problem in our OLSs, i.e., with only age being not binary, we set

$$X = (1, \text{age}, \text{r2}, \text{r3}, \text{r4}, \text{r5}, \text{r6}, \text{r7}, \text{r9}, \text{SMSA}_{66}, \text{SMSA}, \text{South})'.$$

The data set is old, but this suits well our purpose of finding racial discrimination effect on wage, which consists of the direct effect, the indirect effect through college education, and the interaction effect of black and college education. When gender discrimination cases were argued in court, often the counter-argument was that females were less educated/qualified, but lower education/qualification itself might have been due to gender discrimination. Hence, it is important to account for the indirect discrimination through missed education opportunities, but doing so with recent data would be difficult because discrimination due to denied education opportunities is unlikely to be present in recent data. For this reason, using an old data set as ours is advantageous.

Table 3 presents the estimation results with $X = X_a = X_1 = X_d = X_m = X_{dm}$. OLS_{v1} uses X for all unknown $\psi(X)$ and $\mu(X)$ functions in (3.1) and (3.2), whereas OLS_{v2} uses additionally the interaction terms between age and the other elements (r2~South) of X. OLS_{v2} uses almost twice as many covariates as OLS_{v1} does, but both estimates are almost the same in Table 1, and all effects are statistically significant. Table 3 also presents the results for OLS_{v0} that is the usual OLS for constant effect models with "intercept" $X'\beta_x$. OLS_{v0} differs much from OLS_{v1} and OLS_{v2} in terms of the direct (and thus total) effect.

Table 3 further presents GMM_{v1} for the M-CRF that uses the same X as OLS_{v1} does. GMM_{v1} specifies $\psi_1(X) = \Phi(\alpha_c'X)$ and $\psi_d(X) = \Phi(\alpha_t'X) - \Phi(\alpha_c'X)$ to take "$M \in [0,1]$" into account; GMM_{v1} still applies OLS to the Y-CRF. For GMM_{v1}, we used bootstrap to obtain 95% asymptotic confidence intervals (CIs) with 2,000 bootstrap repetitions, and then obtained the ad-hoc "implied Sd" by dividing the CI width by 2 × 1.96 due to "CI width ≈ 2 × 1.96 × Sd." The GMM_{v1} results are almost the same as those of OLS_{v1}. We also tried "GMM_{v2}," but omitted it, as its bootstrap ran into a singularity problem too often.

The total effect of being black on wage is −22%, with the direct effect −24 to −25%, the indirect effect −2.3 to −2.9% through missed college education, and the interaction effect 4.9 to 5.1%. Had it not been for the indirect effect through college education, the wage discrimination would have been lesser by 2.3 to 2.9%, and college education alleviated the racial discrimination by 4.9 to 5.1% through the interaction effect.

7 Conclusions

A treatment D can affect an outcome Y directly, as well as indirectly through a mediator M. D can also interact with M to affect Y. In the literature of mediation analysis, various (two-way, three-way, and four-way) decompositions of the total effect into sub-effects appeared, and recommending one decomposition over the others is difficult or groundless in general.

In this article, based on "mediative principal stratification" classifying the mediator into four potential types (never taker, complier, defier, and always taker), we found that a particular three-way decomposition for direct, indirect, and interaction effects is appropriate in the sense that the three effects are associated with the right types.

We further showed how to identify the three effects, and in the process, we obtained "CRF's" for M that is linear in D, and for Y that is linear in (D, M, DM), despite no explicit linearity assumptions. The CRF's hold for any Y (binary, count, continuous, ...), and (D, M, DM) carry unknown functions of X as the slopes, which are the basis for the desired direct, indirect, and interaction effects. A practical estimation scenario is specifying the unknown functions of X as linear, or better yet, nonlinear ones depending on the form and range of M and Y. Then, OLS or GMM can be used to estimate the CRF's.

Acknowledgement: The author is grateful to the associate editor and two anonymous reviewers for their helpful comments. The author is also grateful to Bora Kim for her help in proof-reading the paper.

Funding information: This research has been supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337766).

Author contributions: The author confirms the sole responsibility for the conception of the study, presented results, and manuscript preparation.

Compliance with ethical standard: No human or animal subject is involved in this research.

Data availability statement: The data used in this article are publicly available at http://davidcard.berkeley.edu/data_sets.html.

References

[1] MacKinnon D, Fairchild A, Fritz M. Mediation analysis. Ann Rev Psychol. 2007;58:593–614. doi: https://doi.org/10.1146/annurev.psych.58.110405.085542.

[2] Pearl J. Causality. 2nd ed. Cambridge: Cambridge University Press; 2009. doi: https://doi.org/10.1017/CBO9780511803161.

[3] Imai K, Keele L, Yamamoto T. Identification, inference, and sensitivity analysis for causal mediation effects. Stat Sci. 2010;25:51–71. doi: https://doi.org/10.1214/10-STS321.

[4] TenHave T, Joffe M. A review of causal estimation of effects in mediation analyses. Stat Methods Med Res. 2012;21:77–107. doi: https://doi.org/10.1177/0962280210391076.

[5] Preacher K. Advances in mediation analysis: a survey and synthesis of new developments. Ann Rev Psychol. 2015;66:825–52. doi: https://doi.org/10.1146/annurev-psych-010814-015258.

[6] VanderWeele T. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press; 2015.

[7] VanderWeele T. Mediation analysis: a practitioner's guide. Ann Rev Public Health. 2016;37:17–32. doi: https://doi.org/10.1146/annurev-publhealth-032315-021402.

[8] Nguyen T, Schmid I, Stuart E. Clarifying causal mediation analysis for the applied researcher: defining effects based on what we want to learn. Psychol Methods. 2021;26:255–71. doi: https://psycnet.apa.org/10.1037/met0000299.

[9] Lee M, Lee S. Review and comparison of treatment effect estimators using propensity and prognostic scores. Int J Biostat. 2022;18:357–80. doi: https://doi.org/10.1515/ijb-2021-0005.

[10] Choi J, Lee M. Overlap weight and propensity score residual for heterogeneous effects: a review with extensions. J Stat Plan Inference. 2023;222:22–37. doi: https://doi.org/10.1016/j.jspi.2022.04.003.

[11] Baron R, Kenny D. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Personality Soc Psychol. 1986;51:1173–82. doi: https://psycnet.apa.org/10.1037/0022-3514.51.6.1173.
[12] Pearl J. Direct and Indirect Effects. San Francisco, CA: Morgan Kaufmann; 2001. p. 411–20.
[13] Robins J. Semantics of causal DAG models and the identification of direct and indirect effects. Highly Structured in Highly structured stochastic systems. Green P, Hjort N, Richardson S, (Eds.) Oxford: Oxford University Press; 2003. p. 70–81. doi: https://doi.org/10.1093/oso/9780198510550.003.0007.
[14] VanderWeele T. A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology. 2013;24:224–32. doi: https://doi.org/10.1097/EDE.0b013e318281a64e.
[15] VanderWeele T. A unification of mediation and interaction: a four-way decomposition. Epidemiology. 2014;25:749–61. doi: https://doi.org/10.1097/EDE.0000000000000121.
[16] Imbens G, Angrist J. Identification and estimation of local average treatment effects. Econometrica. 1994;62:467–75. doi: https://doi.org/10.2307/2951620.
[17] Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. J Amer Stat Assoc. 1996;91:444–55. doi: https://doi.org/10.1080/01621459.1996.10476902.
[18] Frangakis C, Rubin D. Principal stratification in causal inference. Biometrics. 2002;58:21–9. doi: https://doi.org/10.1111/j.0006-341X.2002.00021.x.
[19] VanderWeele T, Tchetgen Tchetgen E. Attributing effects to interactions. Epidemiology. 2014;25:711–22. doi: https://doi.org/10.1097/EDE.0000000000000096.
[20] Choi J, Lee M. Regression discontinuity with multiple running variables allowing partial effects. Political Anal. 2018;26:258–74. doi: https://doi.org/10.1017/pan.2018.13.
[21] Petersen M, Sinisi S, van der Laan M. Estimation of direct causal effects. Epidemiology. 2006;17:276–84. doi: https://doi.org/10.1097/01.ede.0000208475.99429.2d.
[22] Lee M. Simple least squares estimator for treatment effects using propensity score residuals. Biometrika. 2018;105:149–4. doi: https://doi.org/10.1093/biomet/asx062.
[23] Lee M. Instrument residual estimator for any response variable with endogenous binary treatment. J R Stat Soc (Series B). 2021;83:612–35. doi: https://doi.org/10.1111/rssb.12442.
[24] Mao H, Li L. Flexible regression approach to propensity score analysis and its relationship with matching and weighting. Stat Med. 2020;39:2017–34. doi: https://doi.org/10.1002/sim.8526.
[25] Choi J, Lee G, Lee M. Endogenous treatment effect for any response conditional on control propensity score. Stat Probability Letters. 2023;196:109747. doi: https://doi.org/10.1016/j.spl.2022.109747.
[26] Lee G, Choi J, Lee M. Minimally capturing heterogeneous complier effect of endogenous treatment for any outcome variable. J Causal Inference. 2023;11:20220036. doi: https://doi.org/10.1515/jci-2022-0036.
[27] Hansen L. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50:1029–1054. doi: https://doi.org/https://doi.org/10.2307/1912775.
[28] Card D. Using geographic variation in college proximity to estimate the return to schooling. in: Aspects of labor market behavior: essays in honour of John Vanderkamp. Christofides L, Grant E, Swidinsky R. (Eds.), Toronto: University of Toronto Press; 1995. p. 201–22.
[29] Tan Z. Marginal and nested structural models using instrumental variables. J Amer Stat Assoc. 2010;105:157–9. doi: https://doi.org/10.1198/jasa.2009.tm08299.
[30] Wang L, Robins J, Richardson T. On falsification of the binary instrumental variable model. Biometrika. 2017;104:229–36. doi: https://doi.org/10.1093/biomet/asw064.

Appendix

Proof for Proposition 1. Substituting (2.1) into the total effect $E(Y^{1,M^1} - Y^{0,M^0})$ in (1.3) renders

$$E[Y^{10} + (Y^{11} - Y^{10})M^1 - \{Y^{00} + (Y^{01} - Y^{00})M^0\}].$$

The existing decompositions are obtained by rewriting this expression differently. To obtain M3M, put $Y^{10} - Y^{00}$ together and move $(Y^{11} - Y^{10})M^1$ to the last place:

$$\begin{aligned} & E\{Y^{10} - Y^{00} - (Y^{01} - Y^{00})M^0 + (Y^{11} - Y^{10})M^1\} \\ &= E\{Y^{10} - Y^{00} + (Y^{01} - Y^{00})(M^1 - M^0) - (Y^{01} - Y^{00})M^1 + (Y^{11} - Y^{10})M^1\} \\ &= E\{Y^{10} - Y^{00} + (Y^{01} - Y^{00})(M^1 - M^0) + \Delta Y^{\pm}M^1\}. \end{aligned}$$

□

Proof for Proposition 2. Take $E(\cdot|D, X)$ on the observed $M = M^0 + (M^1 - M^0)D$: due to C(a)(i),

$$E(M|D, X) = E(M^0|D, X) + E(M^1 - M^0|D, X)D = \psi_1(X) + \psi_d(X)D.$$

Then, defining $U_m \equiv M - E(M|D, X)$ renders (3.1).

Before we address (3.2), observe that, using $f(\cdot|X)$ to denote densities/probabilities:

$$\begin{aligned} & f(M^0, M^1, Y^{00}, Y^{01}, Y^{10}, Y^{11}, D|X) \\ &= f(M^0, M^1, Y^{00}, Y^{01}, Y^{10}, Y^{11}|D, X) \cdot f(D|X) \\ &= f(M^0, M^1|D, X) \cdot f(Y^{00}, Y^{01}, Y^{10}, Y^{11}|D, X) \cdot f(D|X) \quad \{\text{due to C(b)}\} \\ &= f(M^0, M^1|D, X) \cdot f(Y^{00}, Y^{01}, Y^{10}, Y^{11}|X) \cdot f(D|X) \quad \{\text{due to C(a)(ii)}\} \\ &= f(D, M^0, M^1|X) \cdot f(Y^{00}, Y^{01}, Y^{10}, Y^{11}|X). \end{aligned}$$

The first and last expressions of this display yield

$$\textbf{C(d)} \ : \ (D, M^0, M^1) \perp\!\!\!\perp (Y^{00}, Y^{01}, Y^{10}, Y^{11})|X \quad \{\Rightarrow \quad (D, M) \perp\!\!\!\perp (Y^{00}, Y^{01}, Y^{10}, Y^{11})|X\}.$$

The implication arrow holds as $M = M^0 + (M^1 - M^0)D$ is determined by (M^0, M^1, D). Now, note that the observed Y and $E(Y|D, M, X)$ are, using C(d) just above,

$$\begin{aligned} & Y = (1 - D)(1 - M)Y^{00} + (1 - D)MY^{01} + D(1 - M)Y^{10} + DMY^{11} \\ &= Y^{00} + (Y^{10} - Y^{00})D + (Y^{01} - Y^{00})M + \Delta Y^{\pm}DM; \\ & E(Y|D, M, X) = \mu_1(X) + \mu_d(X)D + \mu_m(X)M + \mu_{dm}(X)DM. \end{aligned}$$

Defining $U_y \equiv Y - E(Y|D, M, X)$ then yields (3.2). □

Proof for Proposition 3. Rewrite the indirect effect $E[E\{(Y^{01} - Y^{00})(M^1 - M^0)|D, X\}|X]$ as

$$\begin{aligned} & E\{E(Y^{01} - Y^{00}|D, X) \quad E(M^1 - M^0|D, X) \quad |X\} \\ &= E\{E(Y^{01} - Y^{00}|X) \cdot E(M^1 - M^0|X)|X\} = \beta_m' X_m \cdot X_\alpha' \alpha_d, \end{aligned}$$

where C(b) and C(a) are used, respectively. It holds that

$$\begin{aligned} & \sqrt{N}(\hat{\beta}_m' \overline{X_m X_\alpha'} \hat{\alpha}_d - \beta_m' \overline{X_m X_\alpha'} \alpha_d) \\ &= \sqrt{N}(\hat{\beta}_m' \overline{X_m X_\alpha'} \hat{\alpha}_d - \beta_m' \overline{X_m X_\alpha'} \hat{\alpha}_d + \beta_m' \overline{X_m X_\alpha'} \hat{\alpha}_d - \beta_m' \overline{X_m X_\alpha'} \alpha_d) \\ &= \alpha_d' \overline{X_\alpha X_m'} \sqrt{N}(\hat{\beta}_m - \beta_m) + \beta_m' \overline{X_m X_\alpha'} \sqrt{N}(\hat{\alpha}_d - \alpha_d) + o_p(1). \end{aligned}$$

This yields the asymptotic distribution for the indirect effect in Proposition 3.

Analogously, for the interaction effect $E\{E(\Delta Y^{\pm}M^1|D, X)|X\}$, we obtain

$$E\{E(\Delta Y^{\pm}|D, X)E(M^1|D, X)|X\} = \mu_{dm}(X) \cdot E(M^1|X) = \beta_{dm}' X_{dm} \cdot X_\alpha'(\alpha_1 + \alpha_d).$$

This yields the asymptotic distribution for the interaction effect in Proposition 3. The total effect part follows from the sum of the three sub-effects. □

GMM estimation Recall the moment condition $E\{\theta(Z;\ a)\} = 0$ in (4.3). Set $X = X_a = X_1 = X_d = X_m = X_{dm}$ to simplify exposition with little loss of generality, as was already mentioned; then, $Z = (M, X', D)'$. The number of the moments is the same as the dimension of $a \equiv (a_c', a_t')'$, which is a "just-, not over-, identified" case. The GMM minimizes

$$\frac{1}{N}\sum_i \theta(Z_i;\ a)' \cdot W_N^{-1} \cdot \frac{1}{N}\sum_i \theta(Z_i;\ a), \quad \text{and} \quad W_N \equiv \frac{1}{N}\sum_i \theta(Z_i;\ a)\theta(Z_i;\ a)', \tag{7.1}$$

with respect to (wrt) a; W_N is to estimate $E\{\theta(Z;\ a)\theta(Z;\ a)'\}$.

With the derivative $\theta_a(Z_i;\ a_0)$ of $\theta(Z_i;\ a)$ wrt a evaluated at a_0 (stacked row-wise for each moment), the GMM is implemented by iterating with

$$a_1 = a_0 - \left\{\sum_i \theta_a(Z_i;\ a_0) W_N^{-1} \sum_i \theta_{a'}(Z_i;\ a_0)\right\}^{-1} \sum_i \theta_a(Z_i;\ a_0) W_N^{-1} \sum_i \theta(Z_i;\ a_0). \tag{7.2}$$

W_N is evaluated at a_0, and a_1 is to be replaced by a_0 at each iteration until $|a_1 - a_0|$ becomes negligibly small or the minimum of (7.1) is attained.

In our just-identified case with $E\{\theta_a(Z;\ a)\}$ invertible, the iteration reduces to

$$a_1 = a_0 - \left\{\sum_i \theta_{a'}(Z_i;\ a_0)\right\}^{-1} \sum_i \theta(Z_i;\ a_0). \tag{7.3}$$

Denoting the GMM estimator for a as $\hat{a}_{gmm}$, with $E^{-1}(\cdot)$ standing for $\{E(\cdot)\}^{-1}$, $\sqrt{N}(\hat{a}_{gmm} - a)$ is asymptotically normal and has the asymptotic variance:

$$\begin{aligned} &[E\{\theta_a(Z;\ a)\} \cdot E^{-1}\{\theta(Z;\ a)\theta(Z;\ a)'\} \cdot E\{\theta_{a'}(Z;\ a)\}]^{-1} \\ &\quad = E^{-1}\{\theta_{a'}(Z;\ a)\} \cdot E\{\theta(Z;\ a)\theta(Z;\ a)'\} \cdot E^{-1}\{\theta_a(Z;\ a)\}. \end{aligned} \tag{7.4}$$

One "dilemma" is that (7.3) follows also from minimizing (7.1) with W_N removed:

$$\frac{1}{N}\sum_i \theta(Z_i;\ a)' \cdot \frac{1}{N}\sum_i \theta(Z_i;\ a). \tag{7.5}$$

The GMM minimizing this is called the "unweighted GMM," compared with the optimal GMM minimizing (7.1). The minimand matters, because it is used in stopping the iteration and picking up the final estimate. We adopt the unweighted GMM in this article, because, despite the theoretical superiority of the optimal GMM in combining the moment conditions, weighting often results in numerical instability in practice, which was also the case in our simulation study, although not reported there. The asymptotic variance of the unweighted GMM is still (7.4), as can be conjectured from (7.3).

When Y is binary (other forms of non-continuous Y can be dealt with analogously), to apply GMM to the Y-CRF, we can set, for some β parameters:

$$\begin{aligned} &\mu_1(X) \equiv E(Y^{00}|X) = \Phi(\beta_{00}'X), \quad \mu_d(X) \equiv E(Y^{10} - Y^{00}|X) = \Phi(\beta_{10}'X) - \Phi(\beta_{00}'X), \\ &\mu_m(X) \equiv E(Y^{01} - Y^{00}|X) = \Phi(\beta_{01}'X) - \Phi(\beta_{00}'X), \\ &\mu_{dm}(X) \equiv E(\Delta Y^{\pm}|X) = \Phi(\beta_{11}'X) - \Phi(\beta_{01}'X) - \Phi(\beta_{10}'X) + \Phi(\beta_{00}'X). \end{aligned}$$

The moment condition for the Y-CRF is $E\{\kappa(H;\beta)\} = 0$, where $H \equiv (Z', Y)'$ and

$$\begin{aligned} \kappa(H;\beta) \equiv [&Y - \Phi(\beta_{00}'X) - \{\Phi(\beta_{10}'X) - \Phi(\beta_{00}'X)\}D - \{\Phi(\beta_{01}'X) - \Phi(\beta_{00}'X)\}M \\ &- \{\Phi(\beta_{11}'X) - \Phi(\beta_{01}'X) - \Phi(\beta_{10}'X) + \Phi(\beta_{00}'X)\}DM] \cdot (X', X'D, X'M, X'DM)'. \end{aligned}$$

The remaining steps are analogous to the aforementioned unweighted GMM steps for the M-CRF.