Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators

Xiang Zhou

Harvard University, Cambridge, Massachusetts, USA

Correspondence
Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138, USA.
Email: xiang_zhou@fas.harvard.edu

Abstract
Causal mediation analysis concerns the pathways through which a treatment affects an outcome. While most of the mediation literature focuses on settings with a single mediator, a flourishing line of research has examined settings involving multiple mediators, under which path-specific effects (PSEs) are often of interest. We consider estimation of PSEs when the treatment effect operates through $K (\geq 1)$ causally ordered, possibly multivariate mediators. In this setting, the PSEs for many causal paths are not nonparametrically identified, and we focus on a set of PSEs that are identified under Pearl's nonparametric structural equation model. These PSEs are defined as contrasts between the expectations of $2^{K+1}$ potential outcomes and identified via what we call the generalized mediation functional (GMF). We introduce an array of regression-imputation, weighting and 'hybrid' estimators, and, in particular, two $K + 2$-robust and locally semiparametric efficient estimators for the GMF. The latter estimators are well suited to the use of data-adaptive methods for estimating their nuisance functions. We establish the rate conditions required of the nuisance functions for semiparametric efficiency. We also discuss how our framework applies to several estimands that may be of particular interest in empirical applications. The proposed estimators are...
illustrated with a simulation study and an empirical example.

**KEYWORDS**
causal inference, mediation, multiple robustness, path-specific effects, semiparametric efficiency

## 1 | INTRODUCTION

Causal mediation analysis aims to disentangle the pathways through which a treatment affects an outcome. While traditional approaches to mediation analysis have relied on linear structural equation models, along with their stringent parametric assumptions, to define and estimate direct and indirect effects (e.g. Baron & Kenny, 1986), a large body of research has emerged within the causal inference literature that disentangles the tasks of definition, identification and estimation in the study of causal mechanisms. Using the potential outcomes framework (Neyman, 1923; Rubin, 1974), this body of research has provided model-free definitions of direct and indirect effects (Pearl, 2001; Robins & Greenland, 1992), established the assumptions needed for nonparametric identification (Hafeman & VanderWeele, 2011; Imai et al., 2010; Pearl, 2001; Petersen et al., 2006; Robins, 2003; Robins & Greenland, 1992; VanderWeele, 2015), and developed an array of imputation, weighting and multiply robust methods for estimation (e.g. Albert, 2012; Goetgeluk et al., 2009; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen & Shpitser, 2012; VanderWeele, 2015; Vansteelandt et al., 2012; Wodtke & Zhou, 2020; Zheng & van der Laan, 2012).

While the bulk of the causal mediation literature focuses on settings with a single mediator (or a set of mediators considered as a whole), a flourishing line of research has studied settings that involve multiple causally dependent mediators, under which a set of path-specific effects (PSEs) are often of interest (Albert & Nelson, 2011; Avin et al., 2005; Daniel et al., 2015; Lin & VanderWeele, 2017; Miles et al., 2017; Miles et al., 2020; Shpitser, 2013; Steen et al., 2017; VanderWeele & Vansteelandt, 2014; VanderWeele et al., 2014; Vansteelandt & Daniel, 2017). In particular, Daniel et al. (2015) demonstrated a large number of ways in which the total effect of a treatment can be decomposed into PSEs, established the assumptions under which a subset of these PSEs are identified, and provided a parametric method for estimating these effects (see also Albert & Nelson, 2011). More recently, for a particular PSE in the case of two causally ordered mediators, Miles et al. (2020) offered an in-depth discussion of alternative estimation methods, and, utilizing the efficient influence function (EIF) of its identification formula, developed a triply robust and locally semiparametric efficient estimator. This estimator, by virtue of its multiple robustness, is well suited to the use of data-adaptive methods for estimating its nuisance functions.

To date, most of the literature on PSEs has focused on the case of two mediators, and it remains underexplored how the estimation methods developed in previous studies, such as those in VanderWeele et al. (2014) and Miles et al. (2020), generalize to the case of $K(\geq 1)$ causally ordered mediators. This article aims to bridge this gap. First, we observe that despite a multitude of ways in which a PSE can be defined for each causal path from the treatment to the outcome, most of these PSEs are not identified under Pearl’s nonparametric structural equation model. This observation leads us to focus on the much smaller set of PSEs that can be nonparametrically identified. These PSEs are defined as contrasts between the expectations of $2^{K+1}$ potential outcomes, which, in turn, are identified through a formula that can be viewed as an extension of Pearl’s (2001) and
Daniel et al.'s (2015) mediation formulae to the case of $K$ causally ordered mediators. Following Tchetgen Tchetgen and Shpitser (2012), we refer to the identification formula for these expected potential outcomes as the *generalized mediation functional* (GMF).

We then show that the GMF can be estimated via an array of regression, weighting and 'hybrid' estimators. More important, building on its EIF, we develop two multiply robust and locally semiparametric efficient estimators for the GMF. Both of these estimators are $K+2$-robust, in the sense that they are consistent provided that one of $K+2$ sets of nuisance functions is correctly specified and consistently estimated. These multiply robust estimators are well suited to the use of data-adaptive methods for estimating the nuisance functions. We establish rate conditions for consistency and semiparametric efficiency when data-adaptive methods and cross-fitting (Chernozhukov et al., 2018; Zheng & van der Laan, 2011) are used to estimate the nuisance functions.

Compared with existing estimators that have been proposed for causal mediation analysis, the methodology proposed in this article is distinct in its generality. In fact, the doubly robust estimator for the mean of an incomplete outcome (Scharfstein et al., 1999), the triply robust estimator developed by Tchetgen Tchetgen and Shpitser (2012) for the mediation functional in the one-mediator setting (see also Zheng & van der Laan, 2012), and the estimator proposed by Miles et al. (2020) for their particular PSE, can all be viewed as special cases of the $K+2$-robust estimators—when $K = 0, 1, 2$ respectively. Yet, our framework also encompasses important estimands for which semiparametric estimators have not been proposed. To demonstrate the generality of our framework, we show how our multiply robust semiparametric estimators apply to several estimands that may be of particular interest in empirical applications, including the natural direct effect (NDE), the natural/total indirect effect (NIE/TIE), the natural path-specific effect (nPSE) and the cumulative path-specific effect (cPSE). In Supplementary Material E, we discuss how our framework can also be employed to estimate noncausal decompositions of between-group disparities that are widely used in social science research (Fortin et al., 2011).

Before proceeding, we note that in a separate strand of literature, the term ‘multiple robustness’ has been used to characterize a class of estimators for the mean of incomplete data that are consistent if one of several working models for the propensity score or one of several working models for the outcome is correctly specified (e.g. Han, 2014; Han & Wang, 2013). In this paper, we use ‘$V$-robustness’ to characterize estimators that require modelling *multiple parts of the observed data likelihood* and are consistent provided that one of $V$ sets of the corresponding models is correctly specified, in keeping with the terminology in the causal mediation literature. This definition of ‘multiple robustness’ does not imply that a ‘$K+2$-robust’ estimator is necessarily more robust than, for example, a ‘$K+1$-robust’ estimator. First, they may correspond to different estimands that require modelling different parts of the likelihood. For example, the doubly robust estimator of the average treatment effect only involves a propensity score model and an outcome model; it is thus less demanding than Tchetgen Tchetgen and Shpitser’s (2012) triply robust estimator of the mediation functional, which involves an additional model for the mediator. Second, for our semiparametric estimators of the GMF, the ‘$K+2$-robustness’ property is not ‘sharp’ because it can be tightened in various special cases. As we demonstrate in Section 4 and Supplementary Material E, such a tightening may result in a lower $V$ (as in the case of NDE, NIE/TIE, nPSE and cPSE), or a higher $V$ (as in the case of noncausal decompositions of between-group disparities).

The rest of the paper is organized as follows. In Section 2, we define the PSEs of interest, lay out their identification assumptions, and introduce the GMF. In Section 3, we introduce a range
of regression-imputation (RI), weighting, 'hybrid' and multiply robust estimators for the GMF, and present several techniques that could be used to improve the finite sample performance of the multiply robust estimators. In Section 4, we discuss how our results apply to a number of special cases such as the NDE, NIE/TIE, nPSE and cPSE. A simulation study and an empirical example are given in Sections 5 and 6 to illustrate the proposed estimators. Proofs of Theorems 1-4 are given in Supplementary Materials A, C and D. Replication data and code for the simulation study and the empirical example are available at https://doi.org/10.7910/DVN/5TBUM3.

2 | NOTATION, DEFINITIONS AND IDENTIFICATION

To ease exposition, we start with the case of two causally ordered mediators before moving onto the general setting of $K$ mediators.

2.1 | The case of two causally ordered mediators

Let $A$ denote a binary treatment, $Y$ an outcome of interest and $X$ a vector of pretreatment covariates. In addition, let $M_1$ and $M_2$ denote two causally ordered mediators, and assume $M_1$ precedes $M_2$. We allow each of these mediators to be multivariate, in which case the causal relationships among the component variables are left unspecified. A directed acyclic graph (DAG) representing the relationships between these variables is given in the top panel of Figure 1. In this DAG, four possible causal paths exist from the treatment to the outcome, as shown in the lower panels:

- (a) $A \rightarrow Y$;
- (b) $A \rightarrow M_2 \rightarrow Y$;
- (c) $A \rightarrow M_1 \rightarrow Y$; and
- (d) $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$.

A formal definition of PSEs requires the potential outcomes notation for both the outcome and the mediators. Specifically, let $Y(a, m_1, m_2)$ denote the potential outcome under treatment status $a$ and mediator values $M_1 = m_1$ and $M_2 = m_2$, $M_2(a, m_1)$ the potential value of the mediator $M_2$ under treatment status $a$ and mediator value $M_1 = m_1$, and $M_1(a)$ the potential value of $M_1$. 

\[ X \rightarrow A \rightarrow M_1 \rightarrow M_2 \rightarrow Y \]

\[ (a) \quad \begin{array}{llll} A & M_1 & M_2 & Y \\ \end{array} \]

\[ (b) \quad \begin{array}{llll} A & M_1 & M_2 & Y \\ \end{array} \]

\[ (c) \quad \begin{array}{llll} A & M_1 & M_2 & Y \\ \end{array} \]

\[ (d) \quad \begin{array}{llll} A & M_1 & M_2 & Y \\ \end{array} \]

FIGURE 1 | Causal relationships with two causally ordered mediators

Note: $A$ denotes the treatment, $Y$ denotes the outcome of interest, $X$ denotes a vector of pretreatment covariates, and $M_1$ and $M_2$ denote two causally ordered mediators.
the mediator $M_1$ under treatment status $a$. This notation allows us to define nested counterfactuals in the form of $Y(a, M_1(a_1), M_2(a_2, M_1(a_{12})))$, where $a$, $a_1$, $a_2$ and $a_{12}$ can each take 0 or 1. For example, $Y(1, M_1(0), M_2(0, M_1(0)))$ represents the potential outcome in the hypothetical scenario where the subject was treated but the mediators $M_1$ and $M_2$ were set to values they would have taken if the subject had not been treated. Furthermore, if we let $Y(a)$ denote the potential outcome when treatment status is set to $a$ and the mediators $M_1$ and $M_2$ take on their ‘natural’ values under treatment status $a$ (i.e. $M_1(a)$ and $M_2(a, M_1(a))$), we have $Y(a) = Y(a, M_1(a), M_2(a, M_1(a)))$ by construction. This is sometimes referred to as the ‘composition’ assumption (VanderWeele, 2009).

Under the above notation, for each of the causal paths shown in Figure 1, its PSE can be defined in eight different ways, depending on the reference levels chosen for $A$ for each of the other three paths (Daniel et al., 2015). For example, the average direct effect of $A$ on $Y$, that is, the portion of the treatment effect that operates through the path $A \rightarrow Y$, can be defined as

$$\tau_{A \rightarrow Y}(a_1, a_2, a_{12}) = \mathbb{E}[Y(1, M_1(a_1), M_2(a_2, M_1(a_{12}))) - Y(0, M_1(a_1), M_2(a_2, M_1(a_{12})))],$$

where $a_1$, $a_2$ and $a_{12}$ can each take 0 or 1. In particular, $\tau_{A \rightarrow Y}(0, 0, 0)$ corresponds to the natural direct effect (NDE; Pearl, 2001) or pure direct effect (PDE; Robins & Greenland, 1992) if the mediators $M_1$ and $M_2$ are considered as a whole. In a similar vein, the PSEs via $A \rightarrow M_2 \rightarrow Y$, $A \rightarrow M_1 \rightarrow Y$ and $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ can be defined as

$$\tau_{A \rightarrow M_2 \rightarrow Y}(a, a_1, a_{12}) = \mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12}))) - Y(a, M_1(a_1), M_2(0, M_1(a_{12})))],$$

$$\tau_{A \rightarrow M_1 \rightarrow Y}(a, a_2, a_{12}) = \mathbb{E}[Y(a, M_1(1), M_2(a_2, M_1(a_{12}))) - Y(a, M_1(0), M_2(a_2, M_1(a_{12})))],$$

$$\tau_{A \rightarrow M_1 \rightarrow M_2 \rightarrow Y}(a, a_1, a_2) = \mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(1))) - Y(a, M_1(a_1), M_2(a_2, M_1(0))].$$

In addition, if we use $A \rightarrow M_1 \rightarrow Y$ to denote the combination of the causal paths $A \rightarrow M_1 \rightarrow Y$ and $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$, the corresponding PSE for this ‘composite path’ can be defined as

$$\tau_{A \rightarrow M_1 \rightarrow Y}(a, a_2) = \mathbb{E}[Y(a, M_1(1), M_2(a_2, M_1(1))) - Y(a, M_1(0), M_2(a_2, M_1(0))].$$

This quantity reflects the portion of the treatment effect that operates through $M_1$, regardless of whether it further operates through $M_2$ or not. In particular, $\tau_{A \rightarrow M_1 \rightarrow Y}(0, 0)$ is often referred to as the natural indirect effect (NIE; Pearl, 2001) or the pure indirect effect (PIE; Robins & Greenland, 1992) for $M_1$, whereas $\tau_{A \rightarrow M_1 \rightarrow Y}(1, 1)$ is sometimes called the total indirect effect (TIE; Robins & Greenland, 1992) for $M_1$. Note, however, that the term NIE has also been used to denote $\tau_{A \rightarrow M_1 \rightarrow Y}(1, 1)$ (e.g. Tchetgen Tchetgen & Shpitser, 2012). To avoid ambiguity, we use NIE and TIE to denote $\tau_{A \rightarrow M_1 \rightarrow Y}(0, 0)$ and $\tau_{A \rightarrow M_1 \rightarrow Y}(1, 1)$ respectively. By definition, these PSEs are identified if the corresponding expected potential outcomes, that is, $\mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12})))$, are identified. Below, we review the assumptions under which these expected potential outcomes are identified from observed data.

Following Pearl (2009), we use a DAG to encode a nonparametric structural equation model (NPSEM) with mutually independent errors. In this framework, the top panel of Figure 1 implies no unobserved confounding for any of the treatment-mediator, treatment-outcome, mediator-mediator and mediator-outcome relationships. Formally, we invoke the following assumptions.
Assumption 1  Consistency of A on $M_1, (A, M_1)$ on $M_2$, and $(A, M_1, M_2)$ on $Y$: For any unit and any $a, m_1, m_2, M_1 = M_1(a)$ if $A = a; M_2 = M_2(a, m_1)$ if $A = a$ and $M_1 = m_1$; and $Y = Y(a, m_1, m_2)$ if $A = a, M_1 = m_1$, and $M_2 = m_2$.

Assumption 2 Conditional independence among treatment and potential outcomes: for any $a, a_1, a_2, m_1, m_1^*, m_2, (M_2(a_1), M_2(a_2, m_1), Y(a, m_1, m_2)) \perp \perp A|X; (M_2(a_2, m_1), Y(a, m_1, m_2)) \perp \perp M_1(a_1)|X, A$, and $Y(a, m_1, m_2) \perp \perp M_2(a_2, m_2^*)|X, A, M_1$.

Assumption 3 Positivity: $p_{A|X}(a|x) > \epsilon > 0$ whenever $p_X(x) > \epsilon > 0$; $p_{A|X, M_1}(a|x, m_1) > \epsilon > 0$ whenever $p_{X, M_1}(x, m_1) > \epsilon > 0$ whenever $p_{X, M, M_1}(x, m_1, m_2) > \epsilon > 0$ whenever $p_{X, M, M_1}(x, m_1, m_2) > \epsilon > 0$, where $p(\cdot)$ denotes a probability density/mass function.

Note that Assumption 2 involves conditional independence relationships between the so-called cross-world counterfactuals, such as $(M_2(a_2, m_1), Y(a, m_1, m_2)) \perp \perp M_1(a_1)|X, A$. This assumption is a direct consequence of Pearl’s NPSEM with mutually independent errors. It implies, but is not implied by, the sequential ignorability assumption that Robins (2003) invokes in interpreting causal diagrams (see Robins & Richardson, 2010 for an in-depth discussion). In addition, we note that Assumption 2 does not rule out all forms of unobserved confounding for the causal effects of $X$ on its descendants. For example, unobserved variables are permitted (although not shown) in Figure 1 that affect both $X$ and $Y$.

Under Assumptions 1–3, it can be shown that $\mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12})))$ is identified if and only if $a_{12} = a_1$ (Albert & Nelson, 2011; Avin et al., 2005; Daniel et al., 2015). Consequently, none of the PSEs for the path $A \rightarrow M_1 \rightarrow Y$ is identified because given $a_{12}$, either $\mathbb{E}[Y(a, M_1(1), M_2(a_2, M_1(a_{12})))$ or $\mathbb{E}[Y(a, M_1(0), M_2(a_2, M_1(a_{12})))$ is unidentified. Similarly, none of the PSEs for the path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ is identified. Interestingly, the PSEs for the composite path $A \rightarrow M_1 \rightarrow Y$ are all identified, even if $a \neq a_2$. These results echo the recanting witness criterion developed by Avin et al. (2005), which implies that the PSE for a (possibly composite) path from $A$ to $Y$ when $A$ is set to 0 (or 1) for all other paths is identified if and only if the path of interest contains no ‘recanting witness’—a variable $W$ that has an additional path to $Y$ that is not contained in the path of interest. Thus the PSE $\tau_{A \rightarrow M_1 \rightarrow Y}(0, 0, 0)$ is not identified because $M_1$ has an additional path to $Y$ ($M_1 \rightarrow M_2 \rightarrow Y$) that is not contained in $A \rightarrow M_1 \rightarrow Y$, but the PSE $\tau_{A \rightarrow M_1 \rightarrow Y}(0, 0, 0)$ is identified because all possible paths from $M_1$ to $Y$ is contained in $A \rightarrow M_1 \rightarrow Y$.

Because $\mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12})))$ is identified if and only if $a_1 = a_{12}$, we restrict our attention to cases where $a_1 = a_{12}$ and use the following notation:

$$\psi_{a_1, a_2, a} \triangleq \mathbb{E}[Y(a, M_1(a_1), M_2(a_2, M_1(a_{12}))].$$

Under Assumptions 1–3, $\psi_{a_1, a_2, a}$ is identified via the following formula:

$$\psi_{a_1, a_2, a} = \int \int \int \mathbb{E}[Y|x, a, m_1, m_2]dP(m_2|x, a_2, m_1)dP(m_1|x, a_1)dP(x).$$

(1)

For a proof of the above formula, see Daniel et al. (2015). Equation (1) can be seen as an extension of Pearl’s (2001) mediation formula to the case of two causally ordered mediators.

It should be noted that Assumptions 1–3 constitute a sufficient set of conditions that allow us to identify $\psi_{a_1, a_2, a}$ for arbitrary combinations of $a_1$, $a_2$ and $a$. For specific combinations of $a_1$, $a_2$ and $a$, Assumption 2 can be relaxed. For example, $\psi_{100}$ is still identified via Equation (1) when
unobserved confounding exists for the $M_2$-$Y$ relationship, and $\psi_{010}$ is still identified via Equation (1) when unobserved confounding exists for the $M_1$-$Y$ relationship (Miles et al., 2020; Shpitser, 2013).

### 2.2 The case of $K(\geq 1)$ causally ordered mediators

We now generalize the preceding results to the setting where the treatment effect of $A$ on $Y$ operates through $K$ causally ordered, possibly multivariate mediators, $M_1, M_2, \ldots, M_K$. We assume that for any $k < k'$, $M_k$ precedes $M_{k'}$, such that no component of $M_{k'}$ causally affects any component of $M_k$. In a DAG that is consistent with this setup, a directed path from the treatment to the outcome can pass through any combination of the $K$ mediators, resulting in $2^K$ possible paths. Among the $2^K$ paths, each can be switched ‘on’ or ‘off’, creating $2^{2^K}$ potential outcomes. Also, for each of the $2^K$ paths, the corresponding PSE can be defined in $2^{2^{2^K}-1}$ different ways, depending on whether each of the other $2^K - 1$ paths is switched ‘on’ or ‘off’. For example, when $K = 3$, for each causal path from $A$ to $Y$, its PSE can be defined in $2^{3^2 - 1} = 128$ different ways.

As we will see, despite the exponential growth of possible causal paths and the double exponential growth of possible PSEs, most of these PSEs are not identified under the assumptions associated with Pearl’s NPSEM. To fix ideas, let an overbar denote a vector of variables, so that $\overline{M_k} = (M_1, M_2, \ldots, M_k)$, and $\overline{a_k} = (a_1, a_2, \ldots, a_k)$, where $\overline{M_l} = \overline{m_l} = \overline{a_l} = \emptyset$ if $l \leq 0$. In addition, let $[K]$ denote the set $\{1, 2, \ldots, K\}$, and let $a_{K+1}$, instead of $a$, denote the treatment status set to the path $A \rightarrow Y$. Assumptions 1–3 can now be generalized as below.

**Assumption 1*.** Consistency: For any unit, $M_k = M_k(a_k, \overline{m}_{k-1})$ if $A = a_k$ and $\overline{M}_{k-1} = \overline{m}_{k-1}$, $\forall k \in [K]$; and $Y = Y(a_{K+1}, \overline{M}_K)$ if $A = a_{K+1}$ and $\overline{M}_K = \overline{m}_K$.

**Assumption 2*.** Conditional independence among treatment and potential outcomes: $(M_1(a_1), M_2(a_2, \overline{m}_1), \ldots, Y(a_{K+1}, \overline{m}_K)) \perp \! \! \! \perp A \mid X$; and $(M_{k+1}(a_{k+1}, \overline{m}_k), \ldots, M_K(a_K, \overline{m}_{K-1}), Y(a_{K+1}, \overline{m}_K)) \perp \! \! \! \perp M_k(a_k, \overline{m}_{k-1}) \mid X, A, \overline{M}_{k-1}$, $\forall k \in [K]$.

**Assumption 3*.** Positivity: $p_{A|X}(a|x) > \epsilon > 0$ whenever $p_X(x) > 0$; $p_{A|X,\overline{m}}(a|x, \overline{m}) > \epsilon > 0$ whenever $p_X, \overline{m}(x, \overline{m}) > 0$, $\forall k \in [K]$.

Before giving the identification results, we introduce the following notational shorthands:

$$\overline{M}_k(\overline{a}_k) \triangleq (\overline{M}_{k-1}(\overline{a}_{k-1}), M_k(a_k, \overline{M}_{k-1}(\overline{a}_{k-1}))), \forall k \in [K],$$

$$\psi_{\overline{a}} \triangleq \mathbb{E}[Y(a_{K+1}, \overline{M}_k(\overline{a}_k))],$$

where $\overline{M}_k(\overline{a}_k)$ is defined iteratively, with the assumption that $\overline{M}_0(\overline{a}_0) = \emptyset$. For example, when $K = 3$,

$$\psi_{\overline{a}} = \mathbb{E}[Y(a_4, M_1(a_1), M_2(a_2, M_1(a_1)), M_3(a_3, M_1(a_1), M_2(a_2, M_1(a_1))))].$$

Theorem 1 states that $\psi_{\overline{a}}$ is identified under Assumptions 1*–3*.

**Theorem 1** Under Assumptions 1*–3*, we have

$$\psi_{\overline{a}} = \int_x \int_{\overline{m}_k} \mathbb{E}[Y|x, a_{K+1}, \overline{m}_K] \left[ \prod_{k=1}^K dP(m_k|x, a_k, \overline{m}_{k-1}) \right] dP(x). \quad (2)$$
The above equation extends Pearl’s (2001) and Daniel et al.’s (2015) mediation formulae to the case of K causally ordered mediators. Following the terminology of Tchetgen Tchetgen and Shpitser (2012), we refer to the right-hand side of Equation (2) as the GMF. Theorem 1 echoes Avin et al.’s (2005) recanting witness criterion: a potential outcome is identified (in expectation) if the value that a mediator $M_k$ takes, that is, $M_k(a_k)$, is carried over to all future mediators. This result leads us to focus on the set of expected potential outcomes and PSEs that are nonparametrically identified. For example, to assess the mediating role of $M_k$, we focus on the composite causal path $A \rightarrow M_k \rightarrow Y$, where, as before, the squiggle arrow encompasses all possible causal paths from $M_k$ to $Y$. An identifiable PSE for this path can be expressed as

$$
\tau_{A \rightarrow M_k \rightarrow Y} = \psi_{\alpha_k-1, \cdot} - \psi_{\alpha_k-1, 0, \cdot}.
$$

where $a_{k+1} \triangleq (a_{k+1}, \ldots, a_{K+1})$. The notation $\psi_\cdot$ makes it clear that the average total effect (ATE) of $A$ on $Y$ can be decomposed into $K + 1$ identifiable PSEs corresponding to $A \rightarrow Y$ and $A \rightarrow M_k \rightarrow Y$ ($k \in [K]$):

$$
\text{ATE} = \psi_Y - \psi_0 = \left(\psi_{\alpha_k-1, 1} - \psi_{\alpha_k-1, 0, 1}\right)_{A \rightarrow Y} + \sum_{k=1}^{K} \left(\psi_{\alpha_k-1, 1} - \psi_{\alpha_k-1, 0, 1}\right)_{A \rightarrow M_k \rightarrow Y}.
$$

(3)

To be sure, Equation (3) is not the only way of decomposing the ATE. Depending on the order in which the paths $A \rightarrow Y$ and $A \rightarrow M_k \rightarrow Y$ ($k \in [K]$) are considered, there are $(K + 1)!$ different ways of decomposing the ATE. In the above decomposition, $\psi_{\alpha_k-1} - \psi_{\alpha_k-1}$ corresponds to the NDE if the mediators $\bar{M}_K$ are considered as a whole.

## 3 | ESTIMATION

In this section, we focus on the estimation of the GMF, that is, the right-hand side of Equation (2). When Assumptions 1*–3* hold, the GMF is equal to the causal parameter $\psi_\cdot$, but otherwise, it is still a well-defined statistical parameter of potential scientific interest. To distinguish it from the causal parameter $\psi_\cdot$, we henceforth denote the GMF by $\theta_\cdot$.

### 3.1 | MLE, regression-imputation, and weighting

Equation (2) suggests that $\theta_\cdot$ can be estimated via maximum likelihood estimation (MLE) (Miles et al., 2017). Specifically, we can fit a parametric model for each $p(m_k \mid X, a_k, \bar{m}_{k-1})$ ($k \in [K]$) and for $E[Y \mid X, a_{K+1}, \bar{m}_K]$, and then estimate the GMF via the following equation:

$$
\hat{\theta}_\cdot^{\text{MLE}} = \mathbb{P}_n \left[ \int \bar{m}_K E[Y \mid X, a_{K+1}, \bar{m}_K] \left( \prod_{k=1}^{K} \hat{p}(m_k \mid X, a_k, \bar{m}_{k-1})d\nu(m_k) \right) \right],
$$

(4)

where $\mathbb{P}_n[\cdot] = n^{-1} \sum[\cdot]$ and $\nu(\cdot)$ is an appropriate dominating measure. This approach works best when the mediators $M_1, M_2, \ldots, M_K$ are all discrete and the covariates $X$ are low dimensional, in which case the working models for $p(m_k \mid X, a_k, \bar{m}_{k-1})$ are simply models for
the conditional probabilities of $M_k$ that can be reliably estimated. When some of the mediators are continuous/multivariate or when the covariates $X$ are high dimensional, estimates of the corresponding conditional density/probability functions can be unstable and sensitive to model misspecification. This problem could be mitigated by imposing highly constrained functional forms on the conditional means of the mediators and the outcome. For example, when $\mathbb{E}[M_k | x, a_k, \overline{M}_{k-1}]$ and $\mathbb{E}[Y | x, a_{K+1}, \overline{M}_K]$ are all assumed to be linear with no higher-order or interaction terms, $\theta_\pi^{mle}$ will reduce to a simple function of regression coefficients (e.g. Alwin & Hauser, 1975). Yet, the assumptions of linearity and additivity are unrealistic in many applications, which may lead to biased estimates of $\theta_\pi$. Below, we describe several imputation- and weighting-based strategies for estimating $\theta_\pi$.

First, we observe that the GMF can be written as

$$
\theta_\pi = \mathbb{E}[M_1 | x, a_1] \ldots \mathbb{E}[M_k | x, a_k, \overline{M}_{k-1}] \mathbb{E}[Y | x, a_{K+1}, \overline{M}_K].
$$

(5)

This expression suggests that $\theta_\pi$ can be estimated via an iterated RI approach (Zhou & Yamamoto, 2020):

1. Estimate $\mu_k(X, \overline{M}_k)$ by fitting a parametric model for the conditional mean of $Y$ given $(X, A, \overline{M}_k)$ and then setting $A = a_{K+1}$ for all units;
2. For $k = K - 1, \ldots, 0$, estimate $\mu_k(X, \overline{M}_k)$ by fitting a parametric model for the conditional mean of $\mu_{k+1}(X, \overline{M}_{k+1})$ and then setting $A = a_{k+1}$ for all units;
3. Estimate $\theta_\pi$ by averaging the fitted values $\hat{\mu}_0(X)$ among all units:

$$
\hat{\theta}_\pi^{ri} = \mathbb{P}_n[\hat{\mu}_0(X)].
$$

(6)

The RI estimator can be seen as an extension of the imputation strategy proposed by Vansteelandt et al. (2012) for estimating the NDE and NIE in the one-mediator setting. Since this approach requires modelling only the conditional means of observed/imputed outcomes given different sets of mediators, it is more flexible to use with continuous/multivariate mediators than MLE. Nonetheless, because $\mu_k(x, \overline{M}_k)$ is estimated iteratively, correct specification of all of the outcome models is required for $\hat{\theta}_\pi^{ri}$ to be consistent. Thus, in practice, when parametric models are used to estimate $\mu_k(x, \overline{M}_k)$, care should be taken to ensure that the outcome models used to estimate these functions are mutually compatible. For example, if $\mu_1(X, M_1)$ follows a linear model that includes $X$ and $X^2$ as predictors, then the model used to estimate $\mu_0(X) = \mathbb{E}[\mu_1(X, M_1) | X, A = a_1]$ should also include $X$ and $X^2$ in the predictor set.

The GMF can also be written as

$$
\theta_\pi = \mathbb{E} \left[ \frac{I(A = a_{K+1})}{p(a_{K+1} | X)} \left( \prod_{k=1}^{K} \frac{p(M_k | x, a_k, \overline{M}_{k-1})}{p(M_k | x, a_{K+1}, \overline{M}_{k-1})} \right) Y \right].
$$
This expression suggests a weighting estimator of $\theta_\pi$:

$$\hat{\theta}_\pi^{w-m} = \Pr_n \left[ \left( \prod_{k=1}^{K} \frac{\hat{p}(M_k|X, a_k, \overline{M}_{k-1})}{\hat{p}(a_{K+1})} \right) Y \right].$$

(7)

This estimator can be seen as an extension of the weighting estimator proposed in VanderWeele et al. (2014) for the case of two mediators. It shares a limitation of $\hat{\theta}_\pi^{\text{mle}}$ in that it requires estimates of the conditional densities/probabilities of the mediators, which tend to be noisy if the mediators are continuous or multivariate. This problem, however, can be sidestepped by recasting the mediator density ratios, via Bayes’ rule, as odds ratios in terms of the treatment variable:

$$\frac{p(M_k|X, a_k, \overline{M}_{k-1})}{p(M_k|X, a_{K+1}, \overline{M}_{k-1})} = \frac{p(a_k|X, \overline{M}_k)/p(a_{K+1}|X, \overline{M}_k)}{p(a_k|X, \overline{M}_{k-1})/p(a_{K+1}|X, \overline{M}_{k-1})}. $$

This observation leads to an alternative weighting estimator based on estimates of the conditional probabilities of treatment given different sets of mediators:

$$\hat{\theta}_\pi^{w-a} = \Pr_n \left[ \left( \prod_{k=1}^{K} \frac{\hat{p}(a_k|X, \overline{M}_k)}{\hat{p}(a_k|X, \overline{M}_{k-1})} \right) Y \right].$$

(8)

In applications where the mediators are continuous/multivariate, $\hat{\theta}_\pi^{w-a}$ should be easier to work with than $\hat{\theta}_\pi^{w-m}$. Yet, the parameters for $p(a|x, \overline{M}_k)$ are not variationally independent across different values of $k$. As in the case of the RI estimator, care should be taken to ensure the compatibility of the models specified for $p(a|x, \overline{M}_k)$ (see Miles et al., 2020 for some practical recommendations).

The RI approach and the weighting approach can be combined to form various ‘hybrid estimators’ of $\theta_\pi$. For example, in the case of $K = 2$, one can use RI to estimate $\mu_2(x, m_1, m_2)$, another RI step to estimate $\mu_1(x, m_1)$, and weighting to estimate $\theta_\pi$, yielding an ‘RI-RI-W’ estimator:

$$\hat{\theta}_\pi^{ri-ri-w} = \Pr_n \left[ \frac{\hat{p}(a_1|X)}{\hat{p}(a_1|X)} \hat{\mu}_1(X, M_1) \right].$$

(9)

One can also use RI to estimate $\mu_2(x, m_1, m_2)$ and then employ appropriate weights to estimate $\theta_\pi$, which leads to an ‘RI-W-W’ estimator:

$$\hat{\theta}_\pi^{ri-w-w} = \Pr_n \left[ \frac{\hat{p}(M_1|X, a_1)}{\hat{p}(a_2|X)} \frac{\hat{p}(M_1|X, a_1)}{\hat{p}(M_1|X, a_2)} \hat{\mu}_2(X, M_1, M_2) \right].$$

(10)

In fact, with $K$ mediators, there are $2^{K+1}$ different ways to combine RI and weighting, each of which involves estimating $K + 1$ nuisance functions, which entail a choice between $p(a|x)$ and $\mu_0(x)$ and a choice between $p(m_k|x, a, \overline{m}_{k-1})$ and $\mu_k(x, \overline{m}_k)$ for each $k \in [K]$ (see Supplementary Material B for detailed expressions of these hybrid estimators in the case of $K = 2$). As with $\hat{\theta}_\pi^{\text{mle}}, \hat{\theta}_\pi^{ri}, \hat{\theta}_\pi^{w-m}$, and $\hat{\theta}_\pi^{w-a}$, each of these hybrid estimators will be consistent only if the corresponding nuisance functions are all correctly specified and consistently estimated. In applications where the pretreatment covariates $X$ and/or the mediators have many components, all of the above estimators will be prone to model misspecification bias.
3.2 Multiply robust and semiparametric efficient estimation

Henceforth, let \( O = (X, A, \overline{M}_K, Y) \) denote the observed data, and \( P_{np} \) a nonparametric model over \( O \) wherein all laws \( P \) satisfy the positivity assumption described in Section 2.2. In addition, define \( \mu_k(X, \overline{M}_k) \) iteratively as in Equation (5):

\[
\mu_k(X, \overline{M}_k) = \mathbb{E}[Y|X, a_{k+1}, \overline{M}_k],
\]

\[
\mu_k(X, \overline{M}_k) = \mathbb{E}[\mu_{k+1}(X, \overline{M}_{k+1})|X, a_{k+1}, \overline{M}_k], \quad k = K - 1, \ldots, 0.
\]

**Theorem 2** The EIF of \( \theta_\pi \) in \( P_{np} \) is given by

\[
\varphi_\pi(O) = \sum_{k=0}^{K+1} \varphi_k(O),
\]

where

\[ \varphi_0(O) = \mu_0(X) - \theta_\pi, \]

\[ \varphi_k(O) = \mathbb{I}(A = a_k) \frac{p(A|X)}{p(a_k|X)} \left( \prod_{j=1}^{k-1} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_j, \overline{M}_{j-1})} \right) (\mu_k(X, \overline{M}_k) - \mu_{k-1}(X, \overline{M}_{k-1})), \quad k \in [K], \]

\[ \varphi_{K+1}(O) = \mathbb{I}(A = a_{K+1}) \frac{p(A|X)}{p(a_{K+1}|X)} \left( \prod_{j=1}^{K} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_j, \overline{M}_{j-1})} \right) (Y - \mu_K(X, \overline{M}_K)). \]

The semiparametric efficiency bound for any regular and asymptotically linear estimator of \( \theta_\pi \) in \( P_{np} \) is therefore \( \mathbb{E}[(\varphi_\pi(O))^2] \).

We now present two estimators of \( \theta_\pi \) based on the EIF. First, consider the factorized likelihood of \( O: p(O) = p(X)p(A|X)\prod_{k=1}^{K} p(M_k|X, A, \overline{M}_{k-1})p(Y|X, A, \overline{M}_K) \). Suppose we have estimated \( K + 2 \) nuisance functions, each of which corresponds to a component of \( p(O) \): \( \hat{\pi}_0(a|x) \) for \( p(a|x) \), \( \hat{f}_k(m_k|x, a, \overline{m}_{k-1}) \) for \( p(m_k|x, a, \overline{m}_{k-1}) \), and \( \hat{\mu}_K(x, \overline{m}_K) \) for \( \mathbb{E}[Y|x, a_{K+1}, \overline{m}_K] \). The GMF can now be estimated as

\[
\hat{\theta}_\pi^{\text{efi}} = \mathbb{P}_n \left[ \mathbb{I}(A = a_{K+1}) \frac{\hat{\pi}_0(a_{K+1}|X)}{\hat{\pi}_0(a_{K+1}|X)} \left( \prod_{j=1}^{K} \frac{\hat{f}_j(M_j|X, a_{K+1}, \overline{M}_{j-1})}{\hat{f}_j(M_j|X, a_{K+1}, \overline{M}_{j-1})} \right) (Y - \hat{\mu}_K(X, \overline{M}_K)) \right. \\
+ \sum_{k=1}^{K} \mathbb{I}(A = a_k) \frac{\hat{\pi}_0(a_k|X)}{\hat{\pi}_0(a_k|X)} \left( \prod_{j=1}^{k-1} \frac{\hat{f}_j(M_j|X, a_k, \overline{M}_{j-1})}{\hat{f}_j(M_j|X, a_k, \overline{M}_{j-1})} \right) (\hat{\mu}_k^{\text{mle}}(X, \overline{M}_k) - \hat{\mu}_{k-1}^{\text{mle}}(X, \overline{M}_{k-1})) \right],
\]

where

\[ \hat{\mu}_k^{\text{mle}}(X, \overline{M}_k) = \hat{\mu}_K(X, \overline{M}_K) \]

and \( \hat{\mu}_k^{\text{mle}}(X, \overline{M}_k) \) is iteratively constructed as

\[
\hat{\mu}_k^{\text{mle}}(X, \overline{M}_k) = \int \hat{\mu}_k^{\text{mle}}(X, \overline{M}_k, m_{k+1}) \hat{f}_{k+1}(m_{k+1}|X, a_{k+1}, \overline{M}_k) d\nu(m_{k+1}), \quad k = K - 1, \ldots, 0.
\]

When \( M_{k+1} \) involves continuous components, Equation (13) can be evaluated via Monte Carlo simulation.
When some of the mediators are continuous/multivariate, it can be difficult to estimate the conditional distributions \( p(m_k|x, a, \bar{m}_{k-1}) \). In such cases, it is often preferable to estimate the mediator density ratios using the corresponding odds ratios of the treatment variable, and estimate the functions \( \mu_k(x, \bar{m}_k) \) using the RI approach. Specifically, suppose we have estimated \( 2(K + 1) \) nuisance functions: \( \hat{r}_0(a|x) \) for \( p(a|x) \), \( \hat{r}_k(a|x, \bar{m}_k) \) for \( p(a|x, \bar{m}_k) \) (\( k \in [K] \)), and \( \hat{\mu}_k(x, \bar{m}_k) \) for \( \mu_k(x, \bar{m}_k) \) (\( k \in \{0, 1, \ldots, K\} \)), where for \( k < K \), \( \mu_k(x, \bar{m}_k) \) is estimated iteratively by fitting a model for the conditional mean of \( \hat{\mu}_{k+1}(X, \bar{M}_{k+1}) \) given \( (X, A, \bar{M}_k) \) and then setting \( A = a_{k+1} \) for all units. The GMF can then be estimated as

\[
\hat{\theta}_e^{\text{eif}_1} = \mathbb{P}_n \left[ \frac{I(A = a_{K+1})}{\hat{r}_0(a_1|X)} \left( \prod_{j=1}^K \frac{\hat{r}_j(a_j|X, \bar{M}_j)}{\hat{r}_{j-1}(a_{j-1}|X, \bar{M}_{j-1})} \right) (Y - \hat{\mu}_K(X, \bar{M}_K)) + \sum_{k=1}^K \frac{I(A = a_k)}{\hat{r}_0(a_1|X)} \left( \prod_{j=1}^{k-1} \frac{\hat{r}_j(a_j|X, \bar{M}_j)}{\hat{r}_{j-1}(a_{j-1}|X, \bar{M}_{j-1})} \right) (\hat{\mu}_k(X, \bar{M}_K) - \hat{\mu}_{k-1}(X, \bar{M}_{k-1})) + \hat{\mu}_0(X) \right].
\]

(14)

The multiple robustness and semiparametric efficiency of \( \hat{\theta}_e^{\text{eif}_1} \) and \( \hat{\theta}_e^{\text{eif}_2} \) are given below.

**Theorem 3** Let \( \eta_1 = \{\pi_0, f_1, \ldots, f_K, \mu_K\} \) denote the \( K + 2 \) nuisance functions involved in \( \hat{\theta}_e^{\text{eif}_1} \), and \( \eta_2 = \{\pi_0, \ldots, \pi_K, \mu_0, \ldots, \mu_K\} \) denote the \( 2(K + 1) \) nuisance functions involved in \( \hat{\theta}_e^{\text{eif}_2} \). Suppose that Assumption 3* (positivity) and suitable regularity conditions for estimating equations (e.g. Newey & McFadden, 1994, p. 2148) hold. In addition, suppose that \( \mu_K(x, \bar{m}_K) \) is bounded over the support of \( (X, \bar{M}_K) \). Then, when the elements of \( \eta_1 \) and \( \eta_2 \) are estimated via parametric models,

1. \( \hat{\theta}_e^{\text{eif}_1} \) is consistent and asymptotically normal (CAN) if \( K + 1 \) of the \( K + 2 \) nuisance functions in \( \eta_1 \) are correctly specified and their parameter estimates are \( \sqrt{n}\)-consistent; it is semiparametric efficient if all of the \( K + 2 \) nuisance functions in \( \eta_1 \) are correctly specified and their parameter estimates are \( \sqrt{n}\)-consistent.

2. \( \hat{\theta}_e^{\text{eif}_2} \) is CAN if \( \exists k \in \{0, \ldots, K + 1\} \), the first \( k \) treatment models \( \pi_0, \ldots, \pi_{k-1} \) and the last \( K + 1 - k \) outcome models \( \mu_k, \ldots, \mu_K \) in \( \eta_2 \) are correctly specified and their parameter estimates are \( \sqrt{n}\)-consistent; it is semiparametric efficient if all of the treatment and outcome models in \( \eta_2 \) are correctly specified and their parameter estimates are \( \sqrt{n}\)-consistent.

Both \( \hat{\theta}_e^{\text{eif}_1} \) and \( \hat{\theta}_e^{\text{eif}_2} \) are \( K + 2 \)-robust in the sense that they are CAN provided that one of \( K + 2 \) sets of nuisance functions is correctly specified and the corresponding parameter estimates are \( \sqrt{n}\)-consistent. Several special cases are worth noting. First, in the degenerate case where \( K = 0 \), it is clear that both \( \hat{\theta}_e^{\text{eif}_1} \) and \( \hat{\theta}_e^{\text{eif}_2} \) reduce to the standard doubly robust estimator for \( \mathbb{E}[Y(a)] \) (Scharfstein et al., 1999). Second, when \( K = 1 \), \( \hat{\theta}_e^{\text{eif}_1} \) coincides with Tchetgen Tchetgen and Shpitser’s (2012) triply robust estimator for \( \mathbb{E}[Y(1, M(0))] \). Finally, when \( K = 2 \), \( \hat{\theta}_e^{\text{eif}_1} \) is identical to Miles et al.’s (2020) estimator for \( \theta_{010} \). For this case, however, Miles et al. provide a slightly weaker condition than that implied by Theorem 3 for \( \hat{\theta}_e^{\text{eif}_1} \) to be CAN. Specifically, they showed that \( \hat{\theta}_e^{\text{eif}_1} \) remains CAN even if both \( f_1 \) and \( \mu_2 \) are misspecified. In Section 4, we show that the conditions
for $\hat{\theta}_{a}^{\text{eif}_1}$ to be CAN can also be relaxed for several particular types of PSEs, including the nPSE, of which $\psi_{010} - \psi_{000}$ is a special case. For the $K = 2$ case, Miles et al. also noted that the mediator density ratios in $\hat{\mu}_{0}^{\text{mle}}(X)$ can be indirectly estimated through models for $\pi_1$ and $\pi_2$. Clearly, this approach will result in $\hat{\theta}_{a}^{\text{eif}_1}$ if the $\mu_k(x, \bar{m}_k)$ functions are in the meanwhile estimated through RI. The $K + 2$-robustness of $\hat{\theta}_{a}^{\text{eif}_1}$ and $\hat{\theta}_{a}^{\text{eif}_2}$, interestingly, resembles the multiple robustness of Bang and Robins’s (2005) estimator for the mean of a potential outcome with time-varying treatments and time-varying confounders (Luedtke et al., 2017; Molina et al., 2017; Rotnitzky et al., 2017).

To gain some intuition as to why $\hat{\theta}_{a}^{\text{eif}_1}$ is $K + 2$-robust, consider cases in which only one nuisance function in $\eta_{1}$ is misspecified. When only $\pi_0$ is misspecified, all terms inside $\mathbb{P}_n[\cdot]$ but $\hat{\mu}_{0}^{\text{mle}}(X)$ will have a zero mean (asymptotically), leaving only $\mathbb{P}_n[\hat{\mu}_{0}^{\text{mle}}(X)]$ (i.e. the MLE estimator (4)), which is consistent because the corresponding nuisance functions $\{f_1, \ldots, f_K, K\}$ are all correctly specified. When only $\mu_k$ is misspecified, all terms involving $\hat{\mu}_k(X, \bar{M}_K)$ and $\hat{\mu}_{k}^{\text{mle}}(X, \bar{M}_k)$ ($k = 0, 1, \ldots, K - 1$) inside $\mathbb{P}_n[\cdot]$ will have a zero mean (asymptotically), leaving only a weighted average of $\hat{\theta}$ (i.e. the weighting estimator (7)), which is consistent because the corresponding nuisance functions $\{\pi_0, f_1, \ldots, f_K\}$ are all correctly specified. Finally, when only $f_{k'}$ is misspecified (for some $k' \in [K]$), it can be shown that all terms involving $\hat{f}_{k'}$ and $\hat{\mu}_{k'}^{\text{mle}}(X, \bar{M}_{k'}) (\forall k < k')$ inside $\mathbb{P}_n[\cdot]$ will have a zero mean (asymptotically), leaving only a weighted average of $\hat{\mu}_{k'}^{\text{mle}}(X, \bar{M}_{k'})$. The latter constitutes a ‘hybrid’ estimator similar to those mentioned in the previous subsection, and it is consistent in this case because its nuisance functions $\{\pi_0, f_1, \ldots, f_{k'-1}, f_{k'+1}, \ldots, f_K, K, K\}$ are all correctly specified.

The $K + 2$-robustness of $\hat{\theta}_{a}^{\text{eif}_1}$ is due to a similar logic to that of $\hat{\theta}_{a}^{\text{eif}_2}$. Yet, different from $\hat{\theta}_{a}^{\text{eif}_1}$, $\hat{\theta}_{a}^{\text{eif}_2}$ involves estimating $2(K + 1)$ nuisance functions, $K + 1$ for the conditional probabilities of treatment and $K + 1$ for the conditional means of observed/imputed outcomes. Also, unlike $\hat{\theta}_{a}^{\text{eif}_1}$, the treatment models involved in $\hat{\theta}_{a}^{\text{eif}_2}$ are not variationally independent; neither are the outcome models. For example, when $MK \perp\!\!\!\!\perp A|X, \bar{M}_{K-1}$, $\pi_K(A|X, \bar{M}_K)$ should be identical to $\pi_{K-1}(A|X, \bar{M}_{K-1})$; similarly, when $MK \perp\!\!\!\!\perp Y|X, A, \bar{M}_{K-1}$, $\mu_K(X, \bar{M}_K)$ should be identical to $\mu_{K-1}(X, \bar{M}_{K-1})$. Thus, in practice, both the treatment and outcome models should be specified in a mutually compatible way, otherwise some of the conditions in Theorem 3 may fail by design.

The local efficiency of $\hat{\theta}_{a}^{\text{eif}_1}$ and $\hat{\theta}_{a}^{\text{eif}_2}$ is due to the fact that both of the EIF-based estimating Equations (12) and (14) have a zero derivative with respect to the nuisance functions at the truth. This property, referred to as ‘Neyman orthogonality’ by Chernozhukov et al. (2018), implies that first step estimation of the nuisance functions has no first-order effect on the influence functions of $\hat{\theta}_{a}^{\text{eif}_1}$ and $\hat{\theta}_{a}^{\text{eif}_2}$. This property suggests that the nuisance functions can be estimated using data-adaptive/machine learning methods or their ensembles. In this case, these estimators will still be consistent as long as the nuisance functions associated with one of the $K+2$ conditions in Theorem 3 are consistently estimated. For $\hat{\theta}_{a}^{\text{eif}_1}$, an added advantage of employing data-adaptive methods to estimate the nuisance functions is that, by exploring a larger space within $P_{np}$, the risk of model incompatibility is reduced.

When data-adaptive/machine learning methods are used to estimate the nuisance functions, it is advisable to use sample splitting to render the empirical process term asymptotically negligible (Chernozhukov et al., 2018; Newey & Robins, 2018; Zheng & van der Laan, 2011). For example, Chernozhukov et al. (2018) suggest the method of ‘cross-fitting’, which involves the following steps: (a) randomly partition the sample $S$ into $J$ folds: $S_1, S_2, \ldots, S_j$; (b) for each $j$, obtain a fold-specific estimate of the target parameter using only data from $S_j$ (‘main sample’), but
with nuisance functions learned from the remainder of the sample (i.e. $S \setminus S_j$; ‘auxiliary sample’); (c) average these fold-specific estimates to form a final estimate of the target parameter.

When cross-fitting is used, $\hat{\theta}_\alpha^{\text{eff}_1}$ and $\hat{\theta}_\alpha^{\text{eff}_2}$ will be semiparametric efficient if the corresponding nuisance function estimates are all consistent and converge at sufficiently fast rates. For example, a sufficient (but not necessary) condition for $\hat{\theta}_\alpha^{\text{eff}_1}$ and $\hat{\theta}_\alpha^{\text{eff}_2}$ to attain the semiparametric efficiency bound is when all of the nuisance function estimates converge at faster-than-$n^{-1/4}$ rates. More precise conditions are given in Theorem 4.

**Theorem 4** Let $\hat{\eta}_1 = \{\hat{\eta}_0, \hat{J}_1, \ldots, \hat{J}_K, \hat{\mu}_K\}$ and $\hat{\eta}_2 = \{\hat{\eta}_0, \ldots, \hat{\eta}_k, \hat{\mu}_0, \ldots, \hat{\mu}_K\}$ denote estimates of the nuisance functions involved in $\hat{\theta}_\alpha^{\text{eff}_1}$ and $\hat{\theta}_\alpha^{\text{eff}_2}$ respectively. Let $r_n(v)$ denote a mapping from a nuisance function estimator to its $L_2(P)$ convergence rate where $P$ represents the true distribution of $O = (X, A, M_K, Y)$. Suppose that Assumption 3* (positivity) holds for both the true distribution $P$ and its estimates implied by $\hat{\eta}_1$ and $\hat{\eta}_2$, and that all other assumptions required for Theorem 3 hold. Then, when the nuisance functions are estimated via data-adaptive methods and cross-fitting,

1. $\hat{\theta}_\alpha^{\text{eff}_1}$ is consistent if $K + 1$ of the $K + 2$ elements in $\hat{\eta}_1$ are consistent in the $L_2$-norm; it is CAN and semiparametric efficient if all elements in $\hat{\eta}_1$ are consistent in the $L_2$-norm and $\sum_{u,v \in \hat{\eta}_1 \cup \eta_\mu} r_n(u) r_n(v) = o(n^{-1/2})$;
2. $\hat{\theta}_\alpha^{\text{eff}_2}$ is consistent if $\exists k \in [0, \ldots, K + 1], \hat{\eta}_0, \ldots, \hat{\eta}_{k-1}, \hat{\mu}_k, \ldots, \hat{\mu}_K$ are all consistent in the $L_2$-norm; it is CAN and semiparametric efficient if all elements in $\hat{\eta}_2$ are consistent in the $L_2$-norm and $\sum_{j=0}^{K=0} r_n(\hat{\eta}_j) r_n(\hat{\mu}_j) = o(n^{-1/2})$.

The multiple robustness result for $\hat{\theta}_\alpha^{\text{eff}_1}$ echoes Theorem 3. Moreover, the first part of Theorem 4 states that $\hat{\theta}_\alpha^{\text{eff}_1}$ is CAN and semiparametric efficient if all nuisance functions in $\eta_1$ are consistently estimated and, for every two nuisance functions in $\eta_1$, the product of their convergence rates is $o(n^{-1/2})$. Thus $\hat{\theta}_\alpha^{\text{eff}_1}$ is CAN and semiparametric efficient if all of the $K + 2$ nuisance function estimates are consistent and converge at faster-than-$n^{-1/4}$ rates, but it will also attain semiparametric efficiency under alternative conditions. For example, when estimates of the treatment and mediator models $\{\hat{\eta}_0, \hat{J}_1, \ldots, \hat{J}_K\}$ all converge to the truth at a rate of $n^{-1/3}$ and estimates of the outcome model $\hat{\mu}_K$ converge to the truth at a rate of $n^{-1/3}$, the product of the convergence rates of any two elements in $\hat{\eta}_1$ is either $O(n^{-1/3}) O(n^{-1/3}) = O(n^{-2/3})$ or $O(n^{-1/3}) O(n^{-1/5}) = O(n^{-8/15})$, both faster than $O(n^{-1/2})$.

The second part of Theorem 4 states that $\hat{\theta}_\alpha^{\text{eff}_2}$ is consistent if there exists a $k$ such that the first $k$ treatment models and the last $K + 1 - k$ outcome models in $\eta_2$ are consistently estimated, echoing Theorem 3. As with $\hat{\theta}_\alpha^{\text{eff}_1}$, $\hat{\theta}_\alpha^{\text{eff}_2}$ will be CAN and semiparametric efficient if all of the required nuisance functions are consistently estimated and converge at faster-than-$n^{-1/4}$ rates. The rate condition $\sum_{j=0}^{K=0} r_n(\hat{\eta}_j) r_n(\hat{\mu}_j) = o(n^{-1/2})$ appears to be weaker than that for $\hat{\theta}_\alpha^{\text{eff}_1}$ as it involves the sum of only $K + 1$, rather than $\left( \binom{K+2}{2} \right)$, product terms. Because the outcome models are estimated iteratively, the convergence rate of $\hat{\mu}_K$ will in general depend on the convergence rates of $\{\hat{\mu}_{k+1}, \ldots, \hat{\mu}_K\}$. That is, if $r_n(\hat{\mu}_{k+1}) = O(n^\delta), r_n(\hat{\mu}_k)$ is unlikely to be faster than $O(n^\delta)$. Nonetheless, $\hat{\theta}_\alpha^{\text{eff}_2}$ will be CAN and semiparametric efficient under relatively weak conditions—for example, when estimates of the treatment models all converge to the truth at a rate of $n^{-1/3}$ and estimates of the outcome models all converge to the truth at a rate of $n^{-1/5}$, in which case $\sum_{j=0}^{K=0} r_n(\hat{\eta}_j) r_n(\hat{\mu}_j) = \sum_{j=0}^{K=0} O(n^{-8/15}) = o(n^{-1/2})$. 
For inference on $\hat{\theta}_1^{eif}$ and $\hat{\theta}_2^{eif}$, a simple variance estimator can be constructed from the empirical analogue of the EIF, that is, $\hat{\text{ EIF}}_n(\bar{O})/n$. However, unlike $\hat{\theta}_1^{eif}$ and $\hat{\theta}_2^{eif}$, this variance estimator is not multiply robust—it will be consistent only if the conditions for semiparametric efficiency in Theorem 3 or Theorem 4 are satisfied. Thus, when the nuisance functions are estimated using parametric models, the variance estimator constructed from the empirical EIF may be inconsistent even when the corresponding estimator for $\theta_1$ is CAN—for example, when only $K + 1$ of the $K + 2$ nuisance functions involved in $\hat{\theta}_2^{eif}$ are correctly specified. In this case, the nonparametric bootstrap is a convenient approach to more robust inference. When the nuisance functions are estimated using data-adaptive/machine learning methods, however, the nonparametric bootstrap is not theoretically justified, and the EIF-based variance estimator may still be preferred.

### 3.3 Multiply robust regression-imputation estimators

Both of the multiply robust estimators described above involve inverse probability weights. When the positivity assumption is nearly violated, the inverse probability weights tend to be highly variable, which may lead to poor finite sample performance (Kang & Schafer, 2007; Petersen et al., 2012). A variety of methods have been proposed to reduce the influence of highly variable weights on doubly robust and multiply robust estimators in similar settings (e.g. Robins et al., 2007; Seaman & Vansteelandt, 2018; Tchetgen Tchetgen & Shpitser, 2012). Among them, a common strategy is to tailor the estimating equation of the outcome model(s) such that the terms involving inverse probability weights will equal zero, leaving only an RI or ‘substitution’ estimator that typically resides in the parameter space of the estimand. Below, we briefly describe how this approach can be adapted to $\hat{\theta}_1^{eif}$ and $\hat{\theta}_2^{eif}$.

Let us start with $\hat{\theta}_2^{eif}$, which can be written as

$$\hat{\theta}_2^{eif} = \mathbb{P}_n \left[ \hat{w}_K(X, A, \bar{M}_K)(Y - \hat{\mu}_K(X, \bar{M}_K)) \right. $$

$$+ \sum_{k=1}^{K} \hat{w}_{k-1}(X, A, \bar{M}_{k-1})(\hat{\mu}_k(X, \bar{M}_K) - \hat{\mu}_{k-1}(X, \bar{M}_{k-1})) $$

$$+ \hat{\mu}_0(X) \] . \quad (15)$$

where $\hat{w}_k(A, X, \bar{M}_K)$ ($0 \leq k \leq K$) are estimates of the corresponding inverse probability weights as displayed in Equation (14). Note that the nuisance functions $\hat{\mu}_k(X, \bar{M}_K)$ ($0 \leq k \leq K$) here are all estimated via the RI approach. When the corresponding outcome models are fitted via generalized linear models (GLMs) with canonical links, one can either (a) fit weighted GLMs (with an intercept term) for $\hat{\mu}_k(X, \bar{M}_K)$ using $\hat{w}_k(A, X, \bar{M}_K)$ as weights, or (b) add the corresponding inverse probability weight as an additional covariate in these regressions (Robins et al., 2007). Either way, the score equations for GLMs will ensure that all terms inside $\mathbb{P}_n[\cdot]$ but $\hat{\mu}_0(X)$ have a sample mean of zero, leaving only $\mathbb{P}_n[\hat{\mu}_0(X)]$, which will reside in the parameter space of $\theta_1$ if the latter equals the range of the GLM specified for $\mu_0(x)$.

Alternatively, one can use the method of targeted maximum likelihood estimation (TMLE; van Der Laan & Rubin, 2006; Zheng & van der Laan, 2012), which, by fitting each of the outcome models in two steps, will also ensure a zero sample mean for all terms inside $\mathbb{P}_n[\cdot]$ but $\hat{\mu}_0(X)$. This
approach does not require the first-step models to be GLM and thus can be used with a wider range of outcome models. In our case, it involves the following steps:

1. For $k = K, \ldots, 0$
   (a) Using $\hat{\mu}_{k+1}^{\text{tmle}}(X, \overline{M}_{k+1})$ (or, in the case $k = K$, the observed outcome $Y$) as the response variable, obtain a first-step RI estimate of $\mu_k(X, \overline{M}_k)$;
   (b) Fit a one-parameter GLM for the conditional mean of $\hat{\mu}_{k+1}^{\text{tmle}}(X, \overline{M}_{k+1})$ (or, in the case $k = K$, the observed outcome $Y$), using $g(\hat{\mu}_k(X, \overline{M}_k))$ as an offset term and $\hat{w}_k(A, X, \overline{M}_k)$ as the only covariate (without an intercept term), and obtain an updated estimate $\hat{\mu}_{k+1}^{\text{tmle}}(X, \overline{M}_k) = g^{-1}(g(\hat{\mu}_k(X, \overline{M}_k)) + \hat{\beta}_k \hat{w}_k(A, X, \overline{M}_k))$, where $g(\cdot)$ is the link function for the GLM and $\hat{\beta}_k$ is the estimated coefficient on $\hat{w}_k(A, X, \overline{M}_k)$;

2. Obtain the final estimate $\hat{\theta}_{\overline{a}}^{\text{tmle}} = \mathbb{P}_n[\hat{\mu}_0^{\text{tmle}}(X)]$.

In the one-mediator case, the above estimator is similar to the TMLE estimator proposed by Zheng and van der Laan (2012) for the NDE, that is, $\psi_{01} - \psi_{00}$. Since Zheng and van der Laan’s estimand is the NDE instead of the mediation functional, their TMLE procedure involves fitting a model for the ‘mediated mean outcome difference’ (p. 6), that is, $\mathbb{E}[\mathbb{E}[Y|X, A = 1, M] - \mathbb{E}[Y|X, A = 0, M]|X, A = 0]$, instead of the conditional mean of the imputed outcome itself, that is, $\mu_0(X)$.

As with the GLM-based adjustments, the TMLE approach also yields an RI estimator that resides in the parameter space of $\theta_{\overline{a}}$ if the latter equals the range of the model specified for $\mu_0(x)$. It should be noted that when data-adaptive methods are used to obtain first-step estimates of the nuisance functions, sample splitting should be employed so that steps 1(a) and steps 1(b) are implemented on different subsamples. In cross-fitting, for example, steps 1(a) should be implemented in the auxiliary sample $(S \setminus S_j)$ and steps 1(b) implemented in the main sample $S_j$. The method of TMLE can also be used to adjust $\hat{\theta}_{\overline{a}}^{\text{eif1}}$, in which case the first step estimates of $\mu_k(X, \overline{M}_k)$ ($0 \leq k \leq K - 1$) are based on Equation (13), and the weights $\hat{w}_k(A, X, \overline{M}_k)$ ($0 \leq k \leq K$) reflect the corresponding terms in Equation (12).

4 | SPECIAL CASES

We have so far considered $\theta_{\overline{a}}$ for the unconstrained case where $a_1, \ldots, a_{K+1}$ can each take 0 or 1. In many applications, the researcher may be interested in particular causal estimands such as the NDE, the NIE/TIE and natural path-specific effects (nPSEs; Daniel et al., 2015). Below, we discuss how the multiply robust semiparametric estimators of $\theta_{\overline{a}}$ apply to these estimands. In addition, we discuss a set of cPSEs that together compose the ATE. In Supplementary Material E, we connect these cPSEs to noncausal decompositions of between-group disparities that are widely used in the social sciences. For illustrative purposes, we focus on estimators based on $\hat{\theta}_{\overline{a}}^{\text{eif2}}$, although similar results hold for those based on $\hat{\theta}_{\overline{a}}^{\text{eif1}}$. Throughout this section, we maintain Assumptions 1*–3* so that $\theta_{\overline{a}} = \psi_{\overline{a}}$.

4.1 | Natural direct effect (NDE)

The NDE measures the effect of switching treatment status from 0 to 1 in a hypothetical world where the mediators $(M_1, \ldots, M_K)$ were all set to values they would have ‘naturally’ taken for
each unit under treatment status $A = 0$. It is thus given by $\psi_{0k+1} - \psi_{0k+1}$. The first row of Figure 2 illustrates the baseline and comparison interventions associated with the NDE for the case of $K = 2$, where the black solid and dashed arrows for $A \rightarrow M_1, A \rightarrow M_2,$ and $A \rightarrow Y$ denote activated ($A = 1$) and unactivated ($A = 0$) paths respectively. A semiparametric efficient estimator for the NDE can be constructed as

$$\hat{\text{NDE}}_{\text{eif}_2} = \hat{\theta}_{0k+1} - \hat{\theta}_{0k+1}.$$  

(16)

If we treat $\bar{M}_K = (M_1, \ldots, M_K)$ as a whole, $\psi_{0k+1} - \psi_{0k+1}$ coincides with the NDE defined in the single mediator setting. In fact, $\hat{\text{NDE}}_{\text{eif}_2}$ is akin to the semiparametric estimator of the NDE given in Zheng and van der Laan (2012). By contrast, if we use $\hat{\theta}_{\text{eif}_1}$ instead of $\hat{\theta}_{\text{eif}_2}$ in Equation (16), we obtain Tchetgen Tchetgen and Shpitser’s (2012) estimator of the NDE.

Setting $a_1 = \ldots a_{K+1} = 0$ in Equation (14), we have

$$\hat{\theta}_{0k+1} = \mathbb{P}_n \left[ \frac{\mathbb{I}(A = 0)}{\hat{f}_0(0|X)} (Y - \hat{\mu}_0(X)) + \hat{\mu}_0(X) \right],$$  

(17)

FIGURE 2 Illustrations of natural direct effect (NDE), natural indirect effect (NIE), total indirect effect (TIE), natural path-specific effect (nPSE), and cumulative path-specific effect (cPSE) in the case of two mediators Note: $A$ denotes the treatment, $Y$ denotes the outcome of interest, and $M_1$ and $M_2$ denote two causally ordered mediators. Solid and dashed arrows for $A \rightarrow M_1, A \rightarrow M_2,$ and $A \rightarrow Y$ denote activated ($A = 1$) and unactivated ($A = 0$) paths, respectively. Grey arrows $M_1 \rightarrow M_2, M_1 \rightarrow Y,$ and $M_2 \rightarrow Y$ signify that the mediators $M_1$ and $M_2$ are not under direct intervention.
where \( \mu_0(X) = \mathbb{E}[Y|X,A = 0] \). Not surprisingly, \( \hat{\theta}^{\text{eff}}_{0,k+1} \) is the standard doubly robust estimator for \( \mathbb{E}[Y(0)] \), which is consistent if either \( \hat{\pi}_0(0|X) \) or \( \hat{\mu}_0(X) \) is consistent. Similarly, by setting \( a_1 = \ldots a_K = 0 \) and \( a_{K+1} = 1 \) in Equation (14), we have

\[
\hat{\theta}^{\text{eff}}_{0,k+1} = \mathbb{P}_n \left[ \frac{\mathbb{E}(A = 1) \hat{\pi}_K(0|X,\overline{M_K})}{\hat{\pi}_K(0|X)} (Y - \hat{\mu}_K(X,\overline{M_K})) + \frac{\mathbb{E}(A = 0) \hat{\pi}_K(1|X,\overline{M_K})}{\hat{\pi}_K(1|X)} (\hat{\mu}_K(X,\overline{M_K}) - \hat{\mu}_{0,K}(X)) + \hat{\mu}_{0,K}(X) \right].
\]

In contrast to the general case where \( \overline{a}_K \) is unconstrained, \( \hat{\theta}^{\text{eff}}_{0,k+1} \) involves estimating only four nuisance functions: \( \pi_0(a|x), \pi_K(a|x,\overline{m}_K), \mu_0,K(x), \) and \( \mu_K(x,\overline{m}_K) \), where \( \mu_K(x,\overline{m}_K) = \mathbb{E}[Y|x,A = 1,\overline{m}_K] \) and \( \mu_0,K(x) = \mathbb{E}[\mu_K(X,\overline{M_K})|x,A = 0] \). Hence \( \mu_0,K(x) \) can be estimated by fitting a model for the conditional mean of \( \hat{\mu}_K(X,\overline{M_K}) \) given \((X,A)\) and then setting \( A = 0 \) for all units. It follows from Theorem 3 that \( \hat{\theta}^{\text{eff}}_{0,k+1} \) is triply robust in that it is consistent if one of the following three conditions holds: (a) \( \hat{\pi}_0 \) and \( \hat{\pi}_K \) are consistent; (b) \( \hat{\pi}_0 \) and \( \hat{\mu}_K \) are consistent; and (c) \( \hat{\mu}_{0,K} \) and \( \hat{\mu}_K \) are consistent. In the meantime, we know that \( \hat{\theta}^{\text{eff}}_{0,k+1} \) is consistent if either \( \hat{\pi}_0 \) or \( \hat{\mu}_0 \) is consistent. By taking the intersection of the multiple robustness conditions for \( \hat{\theta}^{\text{eff}}_{0,k+1} \) and \( \hat{\theta}^{\text{eff}}_{0,k+1} \), we deduce that \( \hat{\text{NDE}}^{\text{eff}}_{0,k+1} \) is also triply robust, as detailed in Corollary 1.

**Corollary 1** Suppose all assumptions required for Theorem 4 hold. When the nuisance functions are estimated via parametric models, \( \hat{\text{NDE}}^{\text{eff}}_{0,k} \) is CAN provided that one of the following three sets of nuisance functions is correctly specified and its parameter estimates are \( \sqrt{n} \)-consistent: \( \{ \pi_0, \pi_K \}, \{ \pi_0, \mu_K \}, \{ \mu_0, \mu_0,K, \mu_K \} \). \( \hat{\text{NDE}}^{\text{eff}}_{0,k+1} \) is semiparametric efficient if all of the above nuisance functions are correctly specified and their parameter estimates \( \sqrt{n} \)-consistent. When the nuisance functions are estimated via data-adaptive methods and cross-fitting, \( \hat{\text{NDE}}^{\text{eff}}_{0,k+1} \) is CAN and semiparametric efficient if all of the nuisance functions are consistently estimated and \( r_n(\hat{\pi}_0) + r_n(\hat{\pi}_K) + r_n(\hat{\mu}_0,K) + r_n(\hat{\mu}_K) = o(n^{-1/2}) \).

### 4.2 Natural and total indirect effects for \( M_1 \)

In Section 2.1, we noted that \( \psi_{100} - \psi_{000} \) and \( \psi_{111} - \psi_{011} \) correspond to the NIE and TIE for the first mediator \( M_1 \) (illustrated in the second and third rows of Figure 2). This correspondence extends naturally to the case of \( K \) mediators, where the NIE and TIE for \( M_1 \) are given by

\[
\text{NIE}_{M_1} = \psi_{1,0} - \psi_{0}, \quad \text{TIE}_{M_1} = \psi_{1,K+1} - \psi_{0,1},
\]

where \( 1_2 = (0, \ldots 0) \) and \( 1_{-2} = (1, \ldots 1) \) are vectors of length \( K \) representing the fact that \( a_2 = \ldots = a_{K+1} = 0 \) in \( \text{NIE}_{M_1} \) and \( a_2 = \ldots = a_{K+1} = 1 \) in \( \text{TIE}_{M_1} \). Since \( \text{TIE}_{M_1} \) can be obtained by switching the 0s and 1s in \( \text{NIE}_{M_1} \) and then flipping the sign, we focus on \( \text{NIE}_{M_1} \) below, noting that analogous results hold for \( \text{TIE}_{M_1} \).

A semiparametric efficient estimator of \( \text{NIE}_{M_1} \) can be constructed as

\[
\hat{\text{NIE}}^{\text{eff}}_{M_1} = \hat{\theta}^{\text{eff}}_{1,0} - \hat{\theta}^{\text{eff}}_{1,K+1}.
\]

As shown previously, \( \hat{\theta}^{\text{eff}}_{1,K+1} \) is given by the doubly robust estimator (17). Setting \( a_1 = 1 \) and \( a_2 = \ldots a_{K+1} = 0 \) in Equation (14), we obtain
\[
\hat{\theta}_{1,0}^{\text{eff}_1} = \mathbb{P}_n \left[ \mathbb{I}(A = 0) \frac{\hat{\pi}_1(1|X, M_1)}{\hat{\pi}_0(1|X)} (Y - \hat{\mu}_1(X, M_1)) + \mathbb{I}(A = 1) \frac{\hat{\pi}_1(1|X, M_1)}{\hat{\pi}_0(1|X)} (\hat{\mu}_1(X, M_1) - \hat{\mu}_{0,1}(X)) + \hat{\mu}_{0,1}(X) \right].
\]

Like \(\hat{\theta}_{1,0}^{\text{eff}_1}\), \(\hat{\theta}_{1,0}^{\text{eff}_2}\), also involves estimating four nuisance functions: \(\pi_0(a|x), \pi_1(a|x, m_1), \mu_{0,1}(x)\), and \(\mu_1(x, m_1)\), where \(\mu_1(x, m_1) = \mathbb{E}[Y|x, A = 0, m_1]\) and \(\mu_{0,1}(x) = \mathbb{E}[\mu_1(X, M_1)|x, A = 1]\). It follows from Theorem 3 that \(\hat{\theta}_{1,0}^{\text{eff}_2}\) is triply robust in that it is consistent if one of the following three conditions holds: (a) \(\hat{\pi}_0\) and \(\hat{\pi}_1\) are consistent; (b) \(\hat{\pi}_0\) and \(\hat{\pi}_1\) are consistent; and (c) \(\hat{\mu}_{0,1}\) and \(\hat{\mu}_1\) are consistent. By taking the intersection of the multiple robustness conditions for \(\hat{\theta}_{1,0}^{\text{eff}_1}\) and \(\hat{\theta}_{1,0}^{\text{eff}_2}\), we deduce that NIE\(_M_k\) is also triply robust, as detailed in Corollary 2.

**Corollary 2** Suppose all assumptions required for Theorem 4 hold. When the nuisance functions are estimated via parametric models, NIE\(_M_k\) is CAN provided that one of the following three sets of nuisance functions is correctly specified and its parameter estimates are \(\sqrt{n}\)-consistent: \{\(\pi_0, \pi_1\), \(\pi_0, \mu_1\), \(\pi_0, \mu_{0,1}, \mu_1\)\}. NIE\(_M_k\) is semiparametric efficient if all of the above nuisance functions are correctly specified and their parameter estimates are \(\sqrt{n}\)-consistent. When the nuisance functions are estimated via data-adaptive methods and cross-fitting, NIE\(_M_k\) is semiparametric efficient if all of the nuisance functions are consistently estimated and \(r_n(\hat{\pi}_0) r_n(\hat{\mu}_{0,1}) + r_n(\hat{\pi}_1) r_n(\hat{\mu}_1) + r_n(\hat{\pi}_0) r_n(\hat{\mu}_0) = o(n^{-1/2})\).

### 4.3 Natural path-specific effects (nPSEs) for \(M_k\) (\(k \geq 2\))

In the same spirit of the NIE for \(M_1\), the natural path-specific effect (nPSE; Daniel et al., 2015) for mediator \(M_k\) (\(k \geq 2\)) is defined as

\[
\text{nPSE}_{M_k} = \psi_{a_{k-1},0_{k+1}} - \psi_{0_{k+1}}.
\]

It can be interpreted as the effect of activating the path \(A \rightarrow M_k \rightarrow Y\) while all other causal paths are 'switched off', as shown in the fourth row of Figure 2. A semiparametric efficient estimator of \(\text{nPSE}_{M_k}\) can be constructed as

\[
\text{nPSE}_{M_k}^{\text{eff}} = \hat{\theta}_{a_{k-1},0_{k+1}}^{\text{eff}_1} - \hat{\theta}_{0_{k+1}}^{\text{eff}_2}.
\]

If, instead, we use \(\hat{\theta}_{0_{k+1}}^{\text{eff}_1}\) in the above equation, the resulting estimator \(\text{nPSE}_{M_k}^{\text{eff}}\) can be seen as Miles et al.’s (2020) estimator of \(\theta_{010} - \theta_{000}\) applied to \(\tilde{M}_1 = (M_1, M_2, \ldots M_{k-1})\) and \(\tilde{M}_2 = M_k\).

Again, \(\hat{\theta}_{0_{k+1}}^{\text{eff}_1}\) is given by the doubly robust estimator (17). Setting \(a_1 = \ldots a_{k-1} = a_{k+1} = \ldots a_{K+1} = 0\) and \(a_k = 1\) in Equation (14), we obtain

\[
\hat{\theta}_{a_{k-1},0_{k+1}}^{\text{eff}_1} = \mathbb{P}_n \left[ \mathbb{I}(A = 0) \frac{\hat{\pi}_{k-1}(0|X, \overline{M}_{k-1})}{\hat{\pi}_0(0|X)} \frac{\hat{\pi}_1(1|X, \overline{M}_k)}{\hat{\pi}_0(1|X)} (Y - \hat{\mu}_k(X, \overline{M}_k)) + \mathbb{I}(A = 1) \frac{\hat{\pi}_{k-1}(0|X, \overline{M}_{k-1})}{\hat{\pi}_0(0|X)} \frac{\hat{\pi}_1(1|X, \overline{M}_k)}{\hat{\pi}_0(1|X)} (\hat{\mu}_k(X, \overline{M}_k) - \hat{\mu}_{k-1,k}(X, \overline{M}_{k-1})) + \mathbb{I}(A = 0) \frac{\hat{\pi}_{k-1,k}(X, \overline{M}_{k-1})}{\hat{\pi}_0(0|X)} (\hat{\mu}_{0,k-1,k}(X) - \hat{\mu}_{0,k-1,k}(X)) + \hat{\mu}_{0,k-1,k}(X) \right].
\]
We can see that $\hat{\theta}_{0_k-1,1}^{eff}$ involves estimating six nuisance functions: $\pi_0(a|x)$, $\pi_{k-1}(a|x, \bar{m}_{k-1})$, $\pi_k(a|x, \bar{m}_k)$, $\mu_{0,k-1,k}(x)$, $\mu_{k-1,k}(x, \bar{m}_{k-1})$, and $\mu_k(x, \bar{m}_k)$, where $\mu_k(X, \bar{M}_k) = \mathbb{E}[Y|X, A = 0, \bar{M}_k]$, $\mu_{0,k-1,k}(X, \bar{M}_{k-1}) = \mathbb{E}[\mu_k(X, \bar{M}_k)|X, A = 1, \bar{M}_{k-1}]$, and $\mu_{0,k-1,k}(X) = \mathbb{E}[\mu_{0,k-1,k}(X, \bar{M}_{k-1})|X, A = 0]$. Hence $\mu_{k-1,k}(x)$ can be estimated by fitting a model for the conditional mean of $\hat{\mu}_k(X, \bar{M}_k)$ given $(X, A, \bar{M}_{k-1})$ and then setting $A = 1$ for all units, and $\mu_{0,k-1,k}(x)$ can be estimated by fitting a model for the conditional mean of $\hat{\mu}_{k-1,k}(X, \bar{M}_{k-1})$ given $(X, A)$ and then setting $A = 0$ for all units. It follows from Theorem 3 that $\hat{\theta}_{0_k-1,1}^{eff}$ is quadruply robust in that it is consistent if one of the following four conditions holds: (a) $\hat{\pi}_0$, $\hat{\pi}_{k-1}$ and $\hat{\pi}_k$ are consistent; (b) $\hat{\pi}_0$, $\hat{\pi}_{k-1}$ and $\hat{\mu}_k$ are consistent; (c) $\hat{\pi}_0$, $\hat{\mu}_{k-1,k}$ and $\hat{\mu}_k$ are consistent; and (d) $\hat{\mu}_{0,k-1,k}$, $\hat{\mu}_{k-1,k}$ and $\hat{\mu}_k$ are consistent. By taking the intersection of the multiple robustness conditions for $\hat{\theta}_{0_k-1,1}^{eff}$ and $\hat{\theta}_{0_k}^{eff}$, we deduce that $\hat{nPSE}_{M_k}$ is also quadruply robust, as detailed in Corollary 3.

**Corollary 3** Suppose all assumptions required for Theorem 4 hold. When the nuisance functions are estimated via parametric models, $\hat{nPSE}_{M_k}$ is CAN provided that one of the following four sets of nuisance functions is correctly specified and its parameter estimates are $\sqrt{n}$-consistent:

- $\{\pi_0, \pi_{k-1}, \pi_k\}$, $\{\pi_0, \pi_{k-1}, \mu_k\}$, $\{\pi_0, \mu_{k-1,k}, \mu_k\}$, $\{\mu_0, \mu_{0,k-1,k}, \mu_{k-1,k}, \mu_k\}$.

$nPSE_{M_k}$ is semiparametric efficient if all of the above nuisance functions are correctly specified and their parameter estimates $\sqrt{n}$-consistent. When the nuisance functions are estimated via data-adaptive methods and cross-fitting, $\hat{nPSE}_{M_k}$ is semiparametric efficient if all of the nuisance functions are consistently estimated and $r_n(\hat{\pi}_0)r_n(\hat{\mu}_{0,k-1,k}) + r_n(\hat{\pi}_{k-1})r_n(\hat{\mu}_{k-1,k}) + r_n(\hat{\pi}_k)r_n(\hat{\mu}_k) + r_n(\hat{\pi}_0)r_n(\hat{\mu}_0) = o(n^{1/2})$.

### 4.4 Cumulative path-specific effects (cPSEs) for $M_k$ ($k \geq 2$)

The NDE, NIE and NPE are all defined as the effect of activating one causal path while keeping all other causal paths ‘switched off’. By contrast, in Equation (3), the ATE is decomposed into $K+1$ components, each of which reflects the cumulative contribution of a specific mediator to the ATE. Specifically, the component $\psi_{0,k-1} - \psi_{0,k+1}$ equals the NDE, the component $\psi_{1,k} - \psi_{0,1,k}$ equals TIE, and the component $\psi_{0,k-1,k} - \psi_{0,0,k}$ gauges the additional contribution of the causal path $A \rightarrow M_k \rightarrow Y$ after the causal paths $A \rightarrow M_{k+1} \rightarrow Y$, … $A \rightarrow M_K \rightarrow Y, A \rightarrow Y$ are ‘switched on’. Such a decomposition will be useful in applications where the investigator aims to partition the ATE into its path-specific components.

We define the cPSE for mediator $M_k$ ($k \geq 2$) as

$$cPSE_{M_k} = \psi_{0,k-1,k} - \psi_{0,0,k+1}.$$  

The last row of Figure 2 gives the baseline and comparison interventions associated with cPSE$_{M_k}$ in the case of $K = 2$. A semiparametric efficient estimator for cPSE$_{M_k}$ can be constructed as

$$\hat{cPSE}_{M_k}^{eff} = \hat{\theta}_{0_k-1,k}^{eff} - \hat{\theta}_{0_k}^{eff}.$$

Setting $a_1 = \ldots a_k = 0$ and $a_{k+1} = \ldots = a_{K+1} = 1$ in Equation (14), we obtain

$$\hat{\theta}_{0_k}^{eff} = \sqrt{n} \left[ \frac{\mathbb{E}(A = 1) \hat{\pi}_k(0|X, \bar{M}_k)}{\hat{\pi}_0(0|X)} \hat{\pi}_k(1|X, \bar{M}_k) (Y - \hat{\mu}_k(X, \bar{M}_k)) + \frac{\mathbb{E}(A = 0)}{\hat{\pi}_0(0|X)} (\hat{\mu}_k(X, \bar{M}_k) - \hat{\mu}_{0,k}(X)) + \hat{\mu}_{0,k}(X) \right],$$

(18)
where \( \mu_k(X, \overline{M}_k) = \mathbb{E}[Y | X, A = 1, \overline{M}_k] \) and \( \mu_{0,k}(X) = \mathbb{E}[\mu_k(X, \overline{M}_k) | X, A = 0] \). It follows from Theorem 3 that \( \hat{\theta}^{\text{eif}}_{0,1:2} \) is triply robust in that it is consistent if one of the following three conditions holds: (a) \( \hat{\pi}_0 \) and \( \hat{\pi}_k \) are consistent; (b) \( \hat{\pi}_0 \) and \( \hat{\mu}_k \) are consistent; and (c) \( \hat{\mu}_{0,k} \) and \( \hat{\mu}_k \) are consistent. By replacing \( \hat{\theta}^{\text{eif}}_{0,1:2} \) with \( \hat{\pi}_0 \) in Equation (18), we obtain a similar expression for \( \hat{\theta}^{\text{eif}}_{0,1:2} \), which is also triply robust in that it is consistent if one of the following three conditions holds: (a) \( \hat{\pi}_0 \) and \( \hat{\pi}_{k-1} \) are consistent; (b) \( \hat{\pi}_0 \) and \( \hat{\mu}_{k-1} \) are consistent; and (c) \( \hat{\mu}_{0,k} \) and \( \hat{\mu}_{k-1} \) are consistent. As a result, \( c\text{PSE}_{M_0}^{\text{eif}} \) involves fitting seven working models—for \( \pi_0(a|x) \), \( \pi_{k-1}(a|x, \overline{m}_{k-1}) \), \( \pi_k(a|x, \overline{m}_k) \), \( \mu_{k-1}(x, \overline{m}_{k-1}) \), \( \mu_0(x, \overline{m}_k) \), \( \mu_k(x, \overline{m}_k) \), and \( \mu_{0,k}(x) \) By taking the product of the above nuisance functions and \( \hat{\mu}_{0,k} \), we deduce that \( c\text{PSE}_{M_0}^{\text{eif}} \) is quintuply robust in that it is consistent if one of five sets of nuisance functions is correctly specified and consistently estimated, as detailed in Corollary 4.

**Corollary 4** Suppose all assumptions required for Theorem 4 hold. When the nuisance functions are estimated via parametric models, \( c\text{PSE}_{M_0}^{\text{eif}} \) is CAN provided that one of the following five sets of nuisance functions is correctly specified and its parameter estimates are \( \sqrt{n} \)-consistent:

\{ \( \pi_0 \), \( \pi_{k-1} \), \( \pi_k \); \( \pi_0 \), \( \pi_{k-1} \), \( \mu_k \); \( \pi_0 \), \( \mu_{k-1} \), \( \pi_k \); \( \mu_0 \), \( \mu_{k-1} \), \( \pi_k \). \( c\text{PSE}_{M_0}^{\text{eif}} \) is semiparametric efficient if all of the above nuisance functions are correctly specified and their parameter estimates \( \sqrt{n} \)-consistent. When the nuisance functions are estimated via data-adaptive methods and cross-fitting, \( c\text{PSE}_{M_0}^{\text{eif}} \) is semiparametric efficient if all of the nuisance functions are consistently estimated and \( r_n(\hat{\pi}_0) r_n(\hat{\mu}_{0,k}) + r_n(\hat{\pi}_0) r_n(\hat{\mu}_{k}) + r_n(\hat{\pi}_{k-1}) r_n(\hat{\mu}_{k-1}) + r_n(\hat{\pi}_k) r_n(\hat{\mu}_k) = o(n^{-1/2}). \)

5 | A SIMULATION STUDY

In this section, we conduct a simulation study to demonstrate the robustness of various estimators under different forms of model misspecification. Specifically, we consider a binary treatment \( A \), a continuous outcome \( Y \), two causally ordered mediators \( M_1 \) and \( M_2 \), and four pretreatment covariates \( X_1, X_2, X_3, X_4 \) generated from the following model:

\[
(U_1, U_2, U_3, U_{XY}) \sim N(0, I_4),
X_j \sim N((U_1, U_2, U_3, U_{XY}) \beta_{Xj}, 1), \quad j = 1, 2, 3, 4,
A \sim \text{Bernoulli}(\text{logit}^{-1}[(1, X_1, X_2, X_3, X_4) \beta_A]),
M_1 \sim N((1, X_1, X_2, X_3, X_4, A) \beta_{M_1}, 1),
M_2 \sim N((1, X_1, X_2, X_3, X_4, A, M_1) \beta_{M_2}, 1),
Y \sim N((1, U_{XY}, X_1, X_2, X_3, X_4, A, M_1, M_2) \beta_Y, 1).
\]

The coefficients \( \beta_{Xj} (1 \leq j \leq 4), \beta_A, \beta_{M_1}, \beta_{M_2}, \beta_Y \) are produced from a set of uniform distributions (see Supplementary Material F for more details). Given the coefficients, we generate 1,000 Monte Carlo samples of size 2,000. Note that in the above model, the unobserved variable \( U_{XY} \) confounds the \( X-Y \) relationship but does not pose an identification threat for \( \psi_m \) and the associated PSEs (i.e. Assumption 2 still holds).

Without loss of generality, we focus on the estimand \( c\text{PSE}_{M_2}^{\text{eif}} \), which we estimate by \( \hat{\theta}_{011} - \hat{\theta}_{001} \). To highlight the general results stated in Theorem 3, we use only estimators for
the generic $\hat{\theta}_a$ (i.e. those described in Section 3). First, we consider the weighting estimator $\hat{\theta}_a^{w-z}$, the RI estimator $\hat{\theta}_a^{ri}$, and the hybrid estimators $\hat{\theta}_a^{ri-w-w}$ and $\hat{\theta}_a^{ri-ri-w}$, where the mediator density ratio involved in $\hat{\theta}_a^{ri-w-w}$ is estimated via the corresponding odds ratio of the treatment variable. We then consider four EIF-based estimators $\hat{\theta}_a^{par,efi}$, $\hat{\theta}_a^{par2,efi}$, $\hat{\theta}_a^{np,efi}$, and $\hat{\theta}_a^{tmle,efi}$. For $\hat{\theta}_a^{par,efi}$ and $\hat{\theta}_a^{par2,efi}$, the nuisance functions are estimated via GLMs. $\hat{\theta}_a^{par,efi}$ differs from $\hat{\theta}_a^{par2,efi}$ in that the outcome models $\mu_2(x, m_1, m_2)$, $\mu_1(x, m_1)$ and $\mu_0(x)$ are fitted using a set of weighted GLMs such that in Equation (15), all terms inside $P_{n}[\cdot]$ but $\tilde{\mu}_0(X)$ have a zero sample mean, yielding an RI estimator that may perform better in finite samples.

All of the above estimators are constructed using estimates of six nuisance functions: $\pi_0(a|x)$, $\pi_1(a|x, m_1)$, $\pi_2(a|x, m_1, m_2)$, $\mu_0(x)$, $\mu_1(x, m_1)$ and $\mu_2(x, m_1, m_2)$. To demonstrate the consequences of model misspecification and the multiple robustness of $\hat{\theta}_a^{par,efi}$ and $\hat{\theta}_a^{par2,efi}$, we generate a set of ‘false covariates’ $Z = (X_1, e^{X_1/2}, (X_3/X_1)^{1/2}, X_4/(e^{X_1/2} + 1))$ and use them to fit a misspecified GLM for each of the nuisance functions (while only the main effects of $Z_1, Z_2, Z_3, Z_4$). We evaluate each of the parametric estimators under five different cases: (a) only $\pi_0, \pi_1, \pi_2$ are correctly specified; (b) only $\pi_0, \pi_1, \mu_2$ are correctly specified; (c) only $\pi_0, \mu_1, \mu_2$ are correctly specified; (d) only $\mu_0, \mu_1, \mu_2$ are correctly specified; and (e) all of the six nuisance functions are misspecified. In theory, $\hat{\theta}_a^{w-z}$ is consistent in case (a), $\hat{\theta}_a^{ri-w-w}$ is consistent in case (b), $\hat{\theta}_a^{ri-ri-w}$ is consistent in case (c), $\hat{\theta}_a^{ri}$ is consistent in case (d), and $\hat{\theta}_a^{par,efi}$ and $\hat{\theta}_a^{par2,efi}$ are consistent in cases (a)–(d). The corresponding estimators of $cPSE_{M_2}$ should follow the same properties.

For the two nonparametric estimators, $\hat{\theta}_a^{np,efi}$ is based on Equation (14), and $\hat{\theta}_a^{tmle,efi}$ is based on the method of TMLE. Like $\hat{\theta}_a^{par,efi}$, $\hat{\theta}_a^{tmle,efi}$ is an RI estimator, which may have better finite-sample performance than $\hat{\theta}_a^{np,efi}$. For both $\hat{\theta}_a^{np,efi}$ and $\hat{\theta}_a^{tmle,efi}$, the nuisance functions are estimated via a super learner (van der Laan et al., 2007) composed of Lasso and random forest, where the feature matrix consists of first-order, second-order and interaction terms of the false covariates $Z$. The super learner is more flexible than a misspecified GLM consisting of only the main effects of $Z$, but it remains agnostic about the true nuisance functions, which are either logit or linear models that depend on $X = (Z_1, 2 \log(Z_2), Z_1Z_3, (1 + e^{Z_1/2})Z_4)$. We obtain nonparametric estimates of $cPSE_{M_2}$ using both fivefold cross-fitting and no cross-fitting.

Results from the simulation study are shown in Figure 3, where each panel corresponds to an estimator, and the y axis is centred at the true value of $cPSE_{M_2}$. The shaded box plots highlight cases under which a given estimator should perform well, and the box plots with a lighter shade in the last two panels denote nonparametric estimators obtained without cross-fitting. From the first four panels, we can see that the weighting, RI and hybrid estimators all behave as expected. They centre around the true value if the requisite nuisance functions are all correctly specified, and deviate from the truth in most other cases. The next four panels show the box plots of the EIF-based estimators. As expected, both of the parametric EIF-based estimators are quadruply robust, as they sampling distributions roughly concentrate around the true value in all of the four cases from (a) to (d). Moreover, it is reassuring to see that when all of the nuisance functions are misspecified (case (e)), the multiply robust estimators do not show a larger amount of bias than those of the other parametric estimators. Finally, both of the nonparametric EIF-based estimators perform reasonably well. When cross-fitting is used, the estimating equation estimator $\hat{cPSE}_{M_2}^{np,efi}$ appears to have a smaller bias than the TMLE estimator $\hat{cPSE}_{M_2}^{tmle,efi}$, but it
FIGURE 3  Sampling distributions of eight different estimators for \( n = 2,000 \). Cases (a)-(e) are described in the main text. The symbols \( y \) and \( n \) denote whether cross-fitting is used to implement the nonparametric estimators \( (y = yes, n = no) \)

occasionally gives rise to extreme estimates. Their 95% Wald confidence intervals, constructed using the estimated variance \( \hat{E}\{[\hat{\phi}_{011} - \hat{\phi}_{001}]^2\}/n \), have close-to-nominal coverage rates—95.5% for \( \hat{\text{cPSE}}_{M_2}^{np, eif_2} \) and 90.9% for \( \hat{\text{cPSE}}_{M_2}^{tmle, eif_2} \). Without cross-fitting, the point estimates exhibit similar distributions, but the coverage rates of the corresponding 95% confidence intervals are somewhat lower—87.3% for \( \hat{\text{cPSE}}_{M_2}^{np, eif_2} \) and 85.8% for \( \hat{\text{cPSE}}_{M_2}^{tmle, eif_2} \).

6 | AN EMPIRICAL APPLICATION

In this section, we illustrate semiparametric estimation of PSEs by analysing the causal pathways through which higher education affects political participation. Prior research suggests that college attendance has a substantial positive effect on political participation in the United States (e.g. Dee, 2004; Milligan et al., 2004). Yet, the mechanisms underlying this causal link remain unclear. The effect of college on political participation may operate through the development of civic and political interest (e.g. Hillygus, 2005), through an increase in economic status (e.g. Kingston et al., 2003), or through other pathways such as social and occupational networks (e.g. Rolfe, 2012). To examine these direct and indirect effects, we consider a causal structure akin to the top panel of Figure 1, where \( A \) denotes college attendance, \( Y \) denotes political participation, and \( M_1 \) and \( M_2 \) denote two causally ordered mediators that reflect (a) economic status, and (b) civic and political interest respectively.

In this model, economic status is allowed to affect civic and political interest but not vice versa, which we consider to be a reasonable approximation to reality. Nonetheless, the conditional independence assumption (Assumption 2) is still strong in this context, as it rules out unobserved confounding for any of the pairwise relationships between college attendance, economic status,
civic and political interest and political participation. Thus, the following analyses should be viewed as an illustration of the proposed methodology rather than a definitive assessment of the PSEs of interest.

We use data from \( n = 2,969 \) individuals in the National Longitudinal Survey of Youth 1997 (NLSY97) who were age 15–17 in 1997 and had completed high school by age 20. The treatment \( A \) is a binary indicator for whether the individual had attended a 2-year or 4-year college by age 20. The outcome \( Y \) is a binary indicator for whether the individual voted in the 2010 general election. We measure economic status (\( M_1 \)) using the respondent’s average annual earnings from 2006 to 2009. To gauge civic and political interest (\( M_2 \)), we use a set of variables that reflect the respondent’s interest in government and public affairs and involvement in volunteering, donation, community group activities between 2007 and 2010. The overlap of the periods in which \( M_1 \) and \( M_2 \) were measured is a limitation of this analysis, and it makes our earlier assumption that \( M_2 \) does not affect \( M_1 \) essential for identifying the PSEs.

To minimize potential bias due to unobserved confounding, we include a rich set of pre-college individual and contextual characteristics in the vector of pretreatment covariates \( X \). They include gender, race, ethnicity, age at 1997, parental education, parental income, parental assets, presence of a father figure, co-residence with both biological parents, percentile score on the Armed Services Vocational Aptitude Battery (ASVAB), high school GPA, an index of substance use (ranging from 0 to 3), an index of delinquency (ranging from 0 to 10), whether the respondent had any children by age 18, college expectation among the respondent’s peers, and a number of school-level characteristics. Descriptive statistics on these pre-college characteristics as well as the mediators and the outcome are given in Supplementary Material G. Some components of \( X \), \( M_1 \) and \( M_2 \) contain a small fraction of missing values. They are imputed via a random-forest-based multiple imputation procedure (with 10 imputed data sets). The standard errors of our parameter estimates are adjusted using Rubin’s (1987) method.

Under Assumptions 1–3 given in Section 2.1, a set of PSEs reflecting the causal paths \( A \rightarrow Y \), \( A \rightarrow M_1 \rightarrow Y \) and \( A \rightarrow M_2 \rightarrow Y \) are identified. For illustrative purposes, we focus on the cPSEs defined in Section 4.4:

\[
\text{ATE} = \psi_{001} - \psi_{000} + \psi_{011} - \psi_{001} + \psi_{111} - \psi_{011}.
\]

Here, the first component is the NDE of college attendance, and the second and third components reflect the amounts of treatment effect that are additionally mediated by civic/political interest and economic status respectively. Since \( M_2 \) is multivariate, it would be difficult to model its conditional distributions directly. We thus estimate the PSEs using the estimator \( \hat{\theta}_{eif}^a_{\psi_{a1},a2,a3} \). Each of the nuisance functions is estimated using a super learner composed with Lasso and random forest. For computational reasons, the feature matrix supplied to the super learner consists of only first-order terms of the corresponding variables. As in our simulation study, we implement two versions of this EIF-based estimator, one based on the original estimating equation \( \hat{\theta}_{np}^a_{\psi_{np}} \), and one based on the method of TMLE \( \hat{\theta}_{tmle}^a_{\psi_{tmle}} \). Fivefold cross-fitting is used to obtain the final estimates.

The results are shown in Table 1. We can see that the two estimators yield similar estimates of the total effect and the PSEs. By \( \hat{\theta}_{np}^a_{\psi_{np}} \), for example, the estimated total effect of college attendance on voting is 0.152, meaning that, on average, college attendance increases the likelihood of
| TABLE 1 | Estimates of total and path-specific effects of college attendance on voting |
|--------------------------------------|-----------------|-----------------|
| **Estimating equation \( \hat{\theta}_{np}^{eif} \)** | **TMLE \( \hat{\theta}_{tmle}^{eif} \)** | |
| Average total effect | 0.152 (0.022) | 0.156 (0.023) |
| Through economic status \( (A \rightarrow M_1 \rightarrow Y) \) | 0.007 (0.005) | 0.002 (0.005) |
| Through civic/political interest \( (A \rightarrow M_2 \rightarrow Y) \) | 0.042 (0.008) | 0.049 (0.008) |
| Direct effect \( (A \rightarrow Y) \) | 0.103 (0.021) | 0.105 (0.021) |

*Note: Numbers in parentheses are estimated standard errors, which are constructed using sample variances of the estimated efficient influence functions and adjusted for multiple imputation via Rubin’s (1987) method.*

Voting in 2010 by about 15 percentage points. The estimated PSE via \( M_2 \) is 0.042, suggesting that a small fraction of the college effect operates through the development of civic and political interest. By contrast, the estimated PSE via economic status is substantively negligible and statistically insignificant. A large portion of the college effect appears to be ‘direct’, that is, operating neither through increased economic status nor through increased civic and political interest.

### 7 CONCLUDING REMARKS

By considering the general case of \( K (\geq 1) \) causally ordered mediators, this paper offers several new insights into the identification and estimation of PSEs. First, under the assumptions associated with Pearl’s NPSEM with mutually independent errors, we have defined a set of PSEs as contrasts between the expectations of \( 2^{K+1} \) potential outcomes, which are identified via the GMF. Second, building on its efficient influence function, we have developed two \( K + 2 \)-robust and semiparametric efficient estimators for the GMF. By virtue of their multiple robustness, these estimators are well suited to the use of data-adaptive methods for estimating their nuisance functions. For such cases, we have established the rate conditions required of the nuisance functions for consistency and semiparametric efficiency.

As we have seen, our proposed methodology is general in that the GMF encompasses a variety of causal estimands such as the NDE, NIE/TIE, nPSE and cPSE. Nonetheless, it does not accommodate PSEs that are not identified under Pearl’s NPSEM, some of which may be scientifically important. For example, social and biomedical scientists are often interested in testing hypotheses about ‘serial mediation’, that is, the degree to which the effect of a treatment operates through multiple mediators sequentially, such as that reflected in the causal path \( A \rightarrow M_1 \rightarrow M_2 \rightarrow Y \) (e.g. Jones et al., 2015). Given that the corresponding PSEs are not nonparametrically identified under Pearl’s NPSEM, previous research has proposed strategies that involve either additional assumptions (Albert & Nelson, 2011) or alternative estimands (Lin & VanderWeele, 2017). We consider semiparametric estimation and inference for these alternative approaches a promising direction for future research.

### ACKNOWLEDGEMENT

The author thanks the Editor, the Associate Editor, two anonymous reviewers, two reviewers from the Alexander and Diviya Magaro Peer Pre-Review Program, and Aleksei Opacic for helpful comments.
DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available at https://urldefense.com/v3/__https://doi.org/10.7910/DVN/5TBUM3__;!!N11eV2iwtsf!vnib2AWtiIMLJXFV0V_9trXnk7phdVZVOGTc_i6lBGDSKo3VuRlkIoZvB4SOZRiwwMRMXjWRyc4ZvMuUKZQS.

ORCID
Xiang Zhou https://orcid.org/0000-0001-8634-7360

REFERENCES
Albert, J.M. (2012) Mediation analysis for nonlinear models with confounding. Epidemiology, 23, 879.
Albert, J.M. & Nelson, S. (2011) Generalized causal mediation analysis. Biometrics, 67, 1028–1038.
Alwin, D.F. & Hauser, R.M. (1975) The decomposition of effects in path analysis. American Sociological Review, 37–47.
Avin, C., Shpitser, I. & Pearl, J. (2005) Identifiability of path-specific effects. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc, pp. 357–363.
Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962–973.
Baron, R.M. & Kenny, D.A. (1986) The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21, C1–C68.
Daniel, R., DeStavola, B., Cousens, S. & Vansteelandt, S. (2015) Causal mediation analysis with multiple mediators. Biometrics, 71, 1–14.
Dee, T.S. (2004) Are there civic returns to education? Journal of Public Economics, 88, 1697–1720.
vander Laan, M.J. & Rubin, D. (2006) Targeted maximum likelihood learning. The International Journal of Biostatistics, 2.
Fortin, N., Lemieux, T. & Firpo, S. (2011) Decomposition methods in economics. In Handbook of labor economics, vol. 4, Amsterdam: Elsevier, pp. 1–102.
Goetgeluk, S., Vansteelandt, S. & Goetzhebeur, E. (2009) Estimation of controlled direct effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70, 1049–1066.
Hafeman, D.M. & VanderWeele, T.J. (2011) Alternative assumptions for the identification of direct and indirect effects. Epidemiology, 753–764.
Han, P. (2014) Multiply robust estimation in regression analysis with missing data. Journal of the American Statistical Association, 109, 1159–1173.
Han, P. & Wang, L. (2013) Estimation with missing data: beyond double robustness. Biometrika, 100, 417–430.
Hillygus, D.S. (2005) The missing link: exploring the relationship between higher education and political engagement. Political Behavior, 27, 25–47.
Imai, K., Keele, L., Yamamoto, T. et al. (2010) Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science, 25, 51–71.
Jones, C.L., Jensen, J.D., Scherr, C.L., Brown, N.R., Christy, K. & Weaver, J. (2015) The health belief model as an explanatory framework in communication research: exploring parallel, serial, and moderated mediation. Health Communication, 30, 566–576.
Kang, J.D. & Schafer, J.L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 22, 523–539.
Kingston, P.W., Hubbard, R., Lapp, B., Schroeder, P. & Wilson, J. (2003) Why education matters. Sociology of Education, 53–70.
vander Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. Statistical Applications in Genetics and Molecular Biology, 6.
Lin, S.-H. & VanderWeele, T. (2017) Interventional approach for path-specific effects. Journal of Causal Inference, 5.
Luedtke, A.R., Sofrygin, O., van der Laan, M.J. & Carone, M. (2017) Sequential double robustness in right-censored longitudinal models. *arXiv preprint arXiv:1705.02459*.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. & Tchetgen Tchetgen, E.J. (2017) Quantifying an adherence path-specific effect of antiretroviral therapy in the nigeria pepfar program. *Journal of the American Statistical Association*, 112, 1443–1452.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. & Tchetgen Tchetgen, E.J. (2020) On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107, 159–172.

Milligan, K., Moretti, E. & Oreopoulos, P. (2004) Does education improve citizenship? Evidence from the united states and the united kingdom. *Journal of Public Economics*, 88, 1667–1695.

Molina, J., Rotnitzky, A., Sued, M. & Robins, J. (2017) Multiple robustness in factorized likelihood models. *Biometrika*, 104, 561–581.

Newey, K. & McFadden, D. (1994) Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 5, 2112–2245.

Newey, W.K. & Robins, J.R. (2018) Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Neyman, J.S. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences*, 10, 1–51.

Pearl, J. (2001) Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc, pp. 411–420.

Pearl, J. (2009) *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

Petersen, M.L., Sinisi, S.E. & van der Laan, M.J. (2006) Estimation of direct causal effects. *Epidemiology*, 17, 276–284.

Petersen, M.L., Porter, K.E., Gruber, S., Wang, Y. & Van Der Laan, M.J. (2012) Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21, 31–54.

Robins, J.M. (2003) Semantics of causal dag models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*, 70–81.

Robins, J.M. & Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–155.

Robins, J.M. & Richardson, T.S. (2010) Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: finding the determinants of disorders and their cures*, 103–158.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007) Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22, 544–559.

Rolfe, M. (2012) *Voter turnout: a social theory of political participation*. Cambridge: Cambridge University Press.

Rotnitzky, A., Robins, J. & Babino, L. (2017) On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. Hoboken: Wiley and Sons.

Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.

Seaman, S.R. & Vansteelandt, S. (2018) Introduction to double robust methods for incomplete data. *Statistical Science*, 33, 184.

Shpitser, I. (2013) Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37, 1011–1035.

Steen, J., Loeys, T., Moerkerke, B. & Vansteelandt, S. (2017) Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186, 184–193.

Tchetgen Tchetgen, E.J. (2013) Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine*, 32, 4567–4580.

Tchetgen Tchetgen, E.J. & Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40, 1816.

VanderWeele, T.J. (2009) Concerning the consistency assumption in causal inference. *Epidemiology*, 20, 880–883.
VanderWeele, T. (2015) *Explanation in causal inference: methods for mediation and interaction.* New York, NY: Oxford University Press.

VanderWeele, T. & Vansteelandt, S. (2014) Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2, 95–115.

VanderWeele, T.J., Vansteelandt, S. & Robins, J.M. (2014) Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25, 300–306.

Vansteelandt, S. & Daniel, R.M. (2017) Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.),* 28, 258.

Vansteelandt, S., Bekaert, M. & Lange, T. (2012) Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1, 131–158.

Wodtke, G.T. & Zhou, X. (2020) Effect decomposition in the presence of treatment-induced confounding: a regression-with-residuals approach. *Epidemiology*, 31, 369–375.

Zheng, W. & van der Laan, M.J. (2011) Cross-validated targeted minimum-loss-based estimation. In *Targeted learning,* New York, NY: Springer, pp. 459–474.

Zheng, W. & van der Laan, M.J. (2012) Targeted maximum likelihood estimation of natural direct effects. *The International Journal of Biostatistics*, 8.

Zhou, X. & Yamamoto, T. (2020) Tracing causal paths from experimental and observational data. *SocArXiv preprint* doi:10.31235/osf.io/2rx6p.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Zhou, X. (2022) Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3), 794–821. Available from: https://doi.org/10.1111/rssb.12487