Semiparametric Estimation for Causal Mediation Analysis with Multiple Causally Ordered Mediators*

Xiang Zhou

Harvard University

November 24, 2020

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 considered settings involving multiple causally ordered mediators, under which a set of path-specific effects (PSEs) are often of interest. We consider estimation of PSEs for the general case where the treatment effect operates through $K(\geq 1)$ causally ordered, possibly multivariate mediators. We first define 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, which are 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 rate conditions required of the nuisance functions for semiparametric efficiency. We also discuss how our framework applies to several causal and noncausal estimands that may be of particular interest in empirical applications. The proposed estimators are illustrated with a simulation study and an empirical example.

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. The author thanks Aleksei Opacic and two reviewers from the Alexander and Diviya Magaro Peer Pre-Review Program for helpful comments.

1 Introduction

Causal mediation analysis aims to disentangle the pathways through which a treatment affects an outcome of interest. 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., Wright 1934; Baron and Kenny 1986), a large body of research has emerged within the causal inference literature that decouples 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 (Robins and Greenland 1992; Pearl 2001), established the assumptions needed for nonparametric identification (Robins and Greenland 1992; Pearl 2001; Robins 2003; Petersen *et al.* 2006; Imai *et al.* 2010; Hafeman and VanderWeele 2011; VanderWeele 2015), and developed an array of imputation, weighting, and multiply robust methods for estimation (e.g., Goetgeluk *et al.* 2009; Albert 2012; Tchetgen Tchetgen and Shpitser 2012; Vansteelandt *et al.* 2012; Zheng and van der Laan 2012; VanderWeele 2015; Wodtke and Zhou 2020).

While the bulk of the causal mediation literature has focused 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 (Avin *et al.* 2005; Albert and Nelson 2011; Shpitser 2013; VanderWeele and Vansteelandt 2014; VanderWeele *et al.* 2014; Daniel *et al.* 2015; Lin and VanderWeele 2017; Miles *et al.* 2017; Steen *et al.* 2017; Vansteelandt and Daniel 2017; Miles *et al.* 2020). In particular, Daniel *et al.* (2015) demonstrated a large number of ways in which the total effect of 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 and 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 of the PSE, 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 seen 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 efficient influence function (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 whenever any of $K+2$ sets of nuisance functions are 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 (Zheng and van der Laan 2011; Chernozhukov *et al.* 2018) are used to estimate the nuisance functions.

Compared with existing estimators that have been proposed for causally 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 and van der Laan 2012), and the estimator proposed by Miles *et al.* (2020) for their particular PSE, can all be seen 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 indirect effect (NIE), the natural path-specific effect (nPSE), and the cumulative path-specific effect (cPSE), and to noncausal decompositions of between-group disparities that are commonly used in the social sciences.

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 and Wang 2013; Han 2014). In this paper, we use "V-robustness" to characterize estimators that require modeling *multiple parts of the observed data likelihood* and are consistent if one of V sets of the corresponding models are 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 modeling 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 will see in Section 4, such a tightening may result in a lower V (as in the case of NDE, NIE, 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, weighting, "hybrid", and multiply robust estimators for the GMF, and present several stabilization techniques that can 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, nPSE, and cPSE. A simulation study and an empirical example are given in Section 5 and Section 6 to illustrate the properties of the proposed estimators. Proofs of Theorems 1-4 are given in Supplementary Materials B-D.

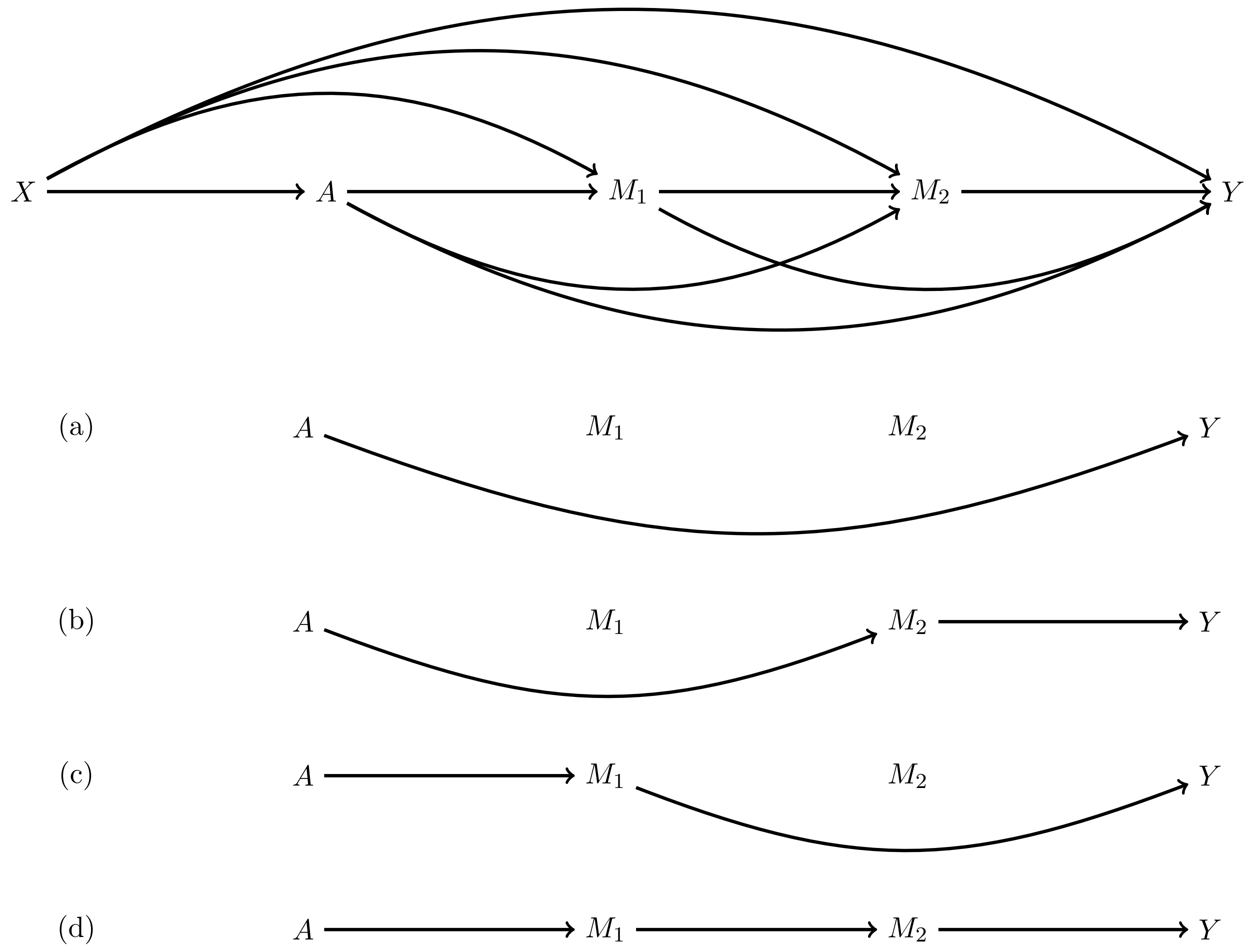

Figure 1: Causal relationships with two causally ordered mediators shown in a DAG.

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.

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) $A \to Y$; (b) $A \to M_2 \to Y$; (c) $A \to M_1 \to Y$; and (d) $A \to M_1 \to M_2 \to Y$.

A formal definition of path-specific effects (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 under treatment status a and mediator value $M_1 = m_1$, and $M_1(a)$ the potential value of the mediator M_1 under treatment status a. This notation allows us to define nested counterfactuals in the form of $Y\big(a, M_1(a_1), M_2(a_2, M_1(a_{12}))\big)$, where a, a_1, a_2, and a_{12} can each take 0 or 1. For example, $Y\big(1, M_1(0), M_2(0, M_1(0))\big)$ 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. Further, 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 $Y(a) = Y\big(a, M_1(a), M_2(a, M_1(a))\big)$ 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, i.e., the portion of the treatment effect that operates through the path $A \to Y$, can be defined as

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

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

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

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

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

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

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

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, $\tau_{A\to M_1\rightsquigarrow Y}(0,0)$ is often referred to as the natural indirect effect (NIE; Pearl 2001) or the pure indirect effect (PIE; Robins and Greenland 1992) with respect to M_1, and $\tau_{A\to M_1\rightsquigarrow Y}(1,1)$ is sometimes called the total indirect effect (TIE; Robins and Greenland 1992) with respect to M_1. By definition, these PSEs are identified if the corresponding expected potential outcomes, i.e., $\mathbb{E}[Y\big(a, M_1(a_1), M_2(a_2, M_1(a_{12}))\big)]$, 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 independent errors. In this framework, the top panel of Figure 1 implies that no unobserved confounding exists 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. *No unmeasured confounding of the A-(M_1, M_2, Y), M_1-(M_2, Y), and M_2-Y relationships: for any $a, a_1, a_2, m_1, m_1^*, m_2$, $\big(M_1(a_1), M_2(a_2, m_1), Y(a, m_1, m_2)\big) \perp\!\!\!\perp A|X$; $\big(M_2(a_2, m_1), Y(a, m_1, m_2)\big) \perp\!\!\!\perp M_1(a_1)|X, A$, and $Y(a, m_1, m_2) \perp\!\!\!\perp M_2(a_2, m_1^*)|X, A, M_1$.*

Assumption 3. *Positivity: $p_{A|X}(a|x) > \epsilon > 0$ whenever $p_X(x) > 0$; $p_{A|X,M_1}(a|x, m_1) > \epsilon > 0$ whenever $p_{X,M_1}(x, m_1) > 0$, and $p_{A|X,M_1,M_2}(a|x, m_1, m_2) > \epsilon > 0$ whenever $p_{X,M,M_2}(x, m_1, m_2) > 0$, where $p(\cdot)$ denotes a probability density/mass function.*

Note that assumption 2 involves conditional independence relationships between so-called cross-world counterfactuals, such as $Y(a, m_1, m_2) \perp\!\!\!\perp M_2(a_2, m_1^*)|X, A, M_1$. This assumption is a direct consequence of Pearl's NPSEM. It implies, but is not implied by, the sequential ignorability assumption that Robins (2003) invokes in interpreting causal diagrams (see Robins and Richardson 2010 for an in-depth discussion). Under assumptions 1-3, it can be shown that $\mathbb{E}[Y\big(a, M_1(a_1), M_2(a_2, M_1(a_{12}))\big)]$ is identified if and only if $a_{12} = a_1$ (Avin *et al.* 2005; Albert and Nelson 2011; Daniel *et al.* 2015). Consequently, none of the PSEs for the path $A \to M_1 \to Y$ is identified because given a_{12}, either $\mathbb{E}[Y\big(a, M_1(1), M_2(a_2, M_1(a_{12}))\big)]$ or $\mathbb{E}[Y\big(a, M_1(0), M_2(a_2, M_1(a_{12}))\big)]$ is unidentified. Similarly, none of the PSEs for the path $A \to M_1 \to M_2 \to Y$ is identified. Interestingly, the PSEs for the composite path $A \to M_1 \rightsquigarrow Y$ 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 $\tau_{A\to M_1\to Y}(0,0,0)$ is not identified because M_1 has an additional path to Y ($M_1 \to M_2 \to Y$) that is not contained in $A \to M_1 \to Y$, but the PSE $\tau_{A\to M_1\rightsquigarrow Y}(0,0)$ is identified because all possible paths from M_1 to Y is contained in $A \to M_1 \rightsquigarrow Y$.

Because $\mathbb{E}[Y\big(a, M_1(a_1), M_2(a_2, M_1(a_{12}))\big)]$ 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

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

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

$$\psi_{a_1,a_2,a} = \iiint \mathbb{E}[Y|x, a, m_1, m_2] dP(m_2|x, a_2, m_1) dP(m_1|x, a_1) dP(x). \tag{1}$$

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, the assumption of no unmeasured confounding can be relaxed. For example, ψ_{100} is still identified via equation (1) when unmeasured confounding exists for the M_2-Y relationship, and ψ_{010} is still identified via equation (1) when unmeasured confounding exists for the M_1-Y relationship (Shpitser 2013; Miles *et al.* 2020).

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, $M_1, M_2, \dots M_K$. 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 go 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^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 $2^{2^3-1} = 128$ 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 $\overline{M}_k = (M_1, M_2, \dots M_k)$, $\overline{m}_k = (m_1, m_2, \dots m_k)$, and $\overline{a}_k = (a_1, a_2, \dots a_k)$, where $\overline{M}_l = \overline{m}_l = \overline{a}_l = \varnothing$ if $l \leq 0$. In addition, let $[K]$ denote the set $\{1, 2, \dots K\}$. 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*. *No unmeasured confounding:* $\big(M_1(a_1), M_2(a_2, \overline{m}_1), \dots Y(a_{K+1}, \overline{m}_K)\big) \perp\!\!\!\perp A|X$*; and* $\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a_{K+1}, \overline{m}_K)\big) \perp\!\!\!\perp M_k(a_k, \overline{m}^*_{k-1})|X, A, \overline{M}_{k-1}, \forall k \in [K]$.

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

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

$$\overline{M}_k(\overline{a}_k) \triangleq \big(\overline{M}_{k-1}(\overline{a}_{k-1}), M_k(a_k, \overline{M}_{k-1}(\overline{a}_{k-1}))\big), \forall k \in [K]$$
$$\psi_{\overline{a}} \triangleq \mathbb{E}[Y(a_{K+1}, \overline{M}_k(\overline{a}_k))],$$

where $\overline{M}_k(\overline{a}_k)$ is defined iteratively, with the assumption that $\overline{M}_0(\overline{a}_0) = \varnothing$. Note that we now use

a_{K+1}, instead of a, to denote the treatment status set to the path $A \to Y$. For example, if $K=3$,

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

Theorem 1 states that $\psi_{\overline{a}}$ is identified under the assumptions of consistency, no unmeasured confounding, and positivity.

Theorem 1. *Under assumptions 1*-3*, we have*

$$\psi_{\overline{a}} = \int_x \int_{\overline{m}_K} \mathbb{E}[Y|x, a_{K+1}, \overline{m}_K]\big[\prod_{k=1}^{K} dP(m_k|x, a_k, \overline{m}_{k-1})\big]dP(x). \tag{2}$$

The above equation extends Pearl's (2001) and Daniel *et al.*'s (2015) mediation formula to the case of K causally ordered mediators. Following the terminology of Tchetgen Tchetgen and Shpitser (2012), we refer to equation (1) as the generalized mediation functional (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, i.e., $M_k(a_k)$, 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 $A \to M_k \rightsquigarrow Y$, 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

$$\tau_{A\to M_k \rightsquigarrow Y}(\overline{a}_{k-1}, \cdot, \underline{a}_{k+1}) = \psi_{\overline{a}_{k-1},1,\underline{a}_{k+1}} - \psi_{\overline{a}_{k-1},0,\underline{a}_{k+1}},$$

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

$$\text{ATE} = \psi_{\overline{1}} - \psi_{\overline{0}} = \underbrace{\psi_{\overline{0}_K,1} - \psi_{\overline{0}_{K+1}}}_{A\to Y} + \sum_{k=1}^{K} \underbrace{\big(\psi_{\overline{0}_{k-1},\underline{1}_k} - \psi_{\overline{0}_k,\underline{1}_{k+1}}\big)}_{A\to M_k \rightsquigarrow Y}. \tag{3}$$

To be sure, equation (3) is not the only way of decomposing $\psi_{\overline{1}} - \psi_{\overline{0}}$. Depending on the order in which the paths $A \to Y$ and $A \to M_k \rightsquigarrow Y$ ($k \in [K]$) are considered, there are $(K+1)!$ different ways of decomposing the ATE. In the above decomposition, $\psi_{\overline{0}_K,1} - \psi_{\overline{0}_{K+1}}$ corresponds to the NDE/PDE if the mediators $\overline{M}_K$ are considered as a whole.

3 Estimation

3.1 MLE, Regression-Imputation, and Weighting

Equation (2) suggests that $\psi_{\overline{a}}$ can be estimated via maximum likelihood (MLE) (Miles *et al.* 2017). Specifically, we can fit a parametric model for each $p(m_k|x, a_k, \overline{m}_{k-1})$ ($k \in [K]$) and for $\mathbb{E}[Y|x, a_{K+1}, \overline{m}_K]$, and then evaluate the GMF via the following equation:

$$\hat{\psi}_{\overline{a}}^{\text{mle}} = \mathbb{P}_n\big[\int_{\overline{m}_K} \hat{\mathbb{E}}[Y|X, a_{K+1}, \overline{m}_K]\big(\prod_{k=1}^{K} \hat{p}(m_k|x, a_k, \overline{m}_{k-1})d\nu(m_k)\big)\big], \tag{4}$$

where $\mathbb{P}_n[\cdot] = n^{-1}\sum_i[\cdot]_i$ 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, in which case the working models for $p(m_k|x, a_k, \overline{m}_{k-1})$ are simply models for the conditional probabilities of M_k. When some of the mediators are continuous/multivariate, estimates of the conditional density/probability functions can be unstable and sensitive to model misspecification. This problem could be circumvented 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, $\hat{\psi}_{\overline{a}}^{\text{mle}}$ will reduce to a simple function of regression coefficients (e.g., Alwin and Hauser 1975). Yet, the assumptions of linearity and additivity are unrealistic in many applications, which may lead to biased estimates of $\psi_{\overline{a}}$. Below, we describe several imputation- and weighting-based strategies for estimating $\psi_{\overline{a}}$.

First, we observe that the GMF can be written as

$$\psi_{\overline{a}} = \mathbb{E}_X\big[\underbrace{\mathbb{E}_{M_1|X,a_1}\ldots\underbrace{\mathbb{E}_{M_K|X,a_K,\overline{M}_{K-1}}\underbrace{\mathbb{E}[Y|X, a_{K+1}, \overline{M}_K]}_{\triangleq\mu_K(X,\overline{M}_K)}}_{\triangleq\mu_{K-1}(X,\overline{M}_{K-1})}}_{\triangleq\mu_0(X)}\big]. \tag{5}$$

This expression suggests that $\psi_{\overline{a}}$ can be estimated via an iterated regression-imputation (RI) approach (Zhou and 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 $\psi_{\overline{a}}$ by averaging the fitted values $\hat{\mu}_0(X)$ among all units:

$$\hat{\psi}_{\overline{a}}^{\text{ri}} = \mathbb{P}_n\big[\hat{\mu}_0(X)\big] \tag{6}$$

The regression-imputation 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 modeling 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, when parametric models are used to estimate the conditional means $\mu_k(x, \overline{m}_k)$, care should be taken to ensure that the outcome models used to estimate these functions are mutually compatible.

The GMF can also be written as

$$\psi_{\overline{a}} = \mathbb{E}\big[\frac{\mathbb{I}(A = a_{K+1})}{p(a_{K+1}|X)}\big(\prod_{k=1}^{K}\frac{p(M_k|X, a_k, \overline{M}_{k-1})}{p(M_k|X, a_{K+1}, \overline{M}_{k-1})}\big)Y\big].$$

This expression points to a weighting estimator of $\psi_{\overline{a}}$:

$$\hat{\psi}_{\overline{a}}^{\text{w-m}} = \mathbb{P}_n\big[\frac{\mathbb{I}(A = a_{K+1})}{\hat{p}(a_{K+1}|X)}\big(\prod_{k=1}^{K}\frac{\hat{p}(M_k|X, a_k, \overline{M}_{k-1})}{\hat{p}(M_k|X, a_{K+1}, \overline{M}_{k-1})}\big)Y\big]. \tag{7}$$

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 $\hat{\psi}_{\overline{a}}^{\text{mle}}$ 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's 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:

$$\hat{\psi}_{\overline{a}}^{\text{w-a}} = \mathbb{P}_n\big[\frac{\mathbb{I}(A = a_{K+1})}{\hat{p}(a_1|X)}\big(\prod_{k=1}^{K}\frac{\hat{p}(a_k|X, \overline{M}_k)}{\hat{p}(a_{k+1}|X, \overline{M}_k)}\big)Y\big]. \tag{8}$$

The estimator $\hat{\psi}_{\overline{a}}^{\text{w-a}}$ is similar to the inverse odds ratio-weighted method proposed by Tchetgen Tchetgen (2013) for estimating the NDE and NIE in the one-mediator setting. In applications where the mediators are continuous/multivariate, $\hat{\psi}_{\overline{a}}^{\text{w-a}}$ should be easier to work with than $\hat{\psi}_{\overline{a}}^{\text{w-m}}$. 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 regression-imputation 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 a detailed discussion).

The regression-imputation approach and the weighting approach can also be combined to form various "hybrid estimators" of $\psi_{\overline{a}}$. For example, in the case of $K = 2$, one can use regression-imputation to estimate $\mu_2(x, m_1, m_2)$, another regression-imputation step to estimate $\mu_1(x, m_1)$, and weighting to estimate $\psi_{\overline{a}_3}$, yielding an "RI-RI-W" estimator:

$$\psi_{\overline{a}_3}^{\text{ri-ri-w}} = \mathbb{P}_n\big[\frac{\mathbb{I}(A = a_1)}{\hat{p}(a_1|X)}\hat{\mu}_1(X, M_1)\big]. \tag{9}$$

Or, one can use regression-imputation to estimate $\mu_2(x, m_1, m_2)$ and then employ appropriate weights to estimate $\psi_{\overline{a}_3}$, which leads to an "RI-W-W" estimator:

$$\psi_{\overline{a}_3}^{\text{ri-w-w}} = \mathbb{P}_n\big[\frac{\mathbb{I}(A = a_2)}{\hat{p}(a_2|X)}\frac{\hat{p}(M_1|X, a_1)}{\hat{p}(M_1|X, a_2)}\hat{\mu}_2(X, M_1, M_2)\big]. \tag{10}$$

In fact, with K mediators, there are 2^{K+1} different ways to combine regression-imputation 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 A for detailed expressions of these hybrid estimators in the case of $K = 2$). As with $\hat{\psi}_{\overline{a}}^{\text{mle}}$, $\hat{\psi}_{\overline{a}}^{\text{ri}}$, $\hat{\psi}_{\overline{a}}^{\text{w-m}}$, and $\hat{\psi}_{\overline{a}}^{\text{w-a}}$, 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 $\mathcal{P}_{\text{np}}$ 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)

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

Theorem 2. *The efficient influence function (EIF) of $\psi_{\overline{a}}$ in $\mathcal{P}_{\rm np}$ is given by*

$$\varphi_{\overline{a}}(O) = \sum_{k=0}^{K+1} \varphi_k(O), \tag{11}$$

where

$$\varphi_0(O) = \mu_0(X) - \psi_{\overline{a}},$$
$$\varphi_k(O) = \frac{\mathbb{I}(A = a_k)}{p(a_k|X)} \Big(\prod_{j=1}^{k-1} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_k, \overline{M}_{j-1})} \Big) \big(\mu_k(X, \overline{M}_k) - \mu_{k-1}(X, \overline{M}_{k-1}) \big), \quad k \in [K],$$
$$\varphi_{K+1}(O) = \frac{\mathbb{I}(A = a_{K+1})}{p(a_{K+1}|X)} \Big(\prod_{j=1}^{K} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_{K+1}, \overline{M}_{j-1})} \Big) \big(Y - \mu_K(X, \overline{M}_K) \big).$$

The semiparametric efficiency bound for any regular and asymptotically linear estimator of $\psi_{\overline{a}}$ in $\mathcal{P}_{\rm np}$ is therefore $\mathbb{E}\big[\big(\varphi_{\overline{a}}(O)\big)^2\big]$.

We now present two estimators of $\psi_{a_1,a_2,a}$ based on the EIF. First, consider the factorized likelihood of O: $p(O) = p(X)p(A|X)\Big(\prod_{k=1}^{K} p(M_k|X, A, \overline{M}_{k-1})\Big)p(Y|X, A, \overline{M}_K)$. 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

$$\begin{aligned}\hat{\psi}_{\overline{a}}^{\rm eif_1} =& \mathbb{P}_n\Big[\frac{\mathbb{I}(A = a_{K+1})}{\hat{\pi}_0(a_{K+1}|X)} \Big(\prod_{j=1}^{K} \frac{\hat{f}_j(M_j|X, a_j, \overline{M}_{j-1})}{\hat{f}_j(M_j|X, a_{K+1}, \overline{M}_{j-1})} \Big) \big(Y - \hat{\mu}_K(X, \overline{M}_K) \big) \\ &+ \sum_{k=1}^{K} \frac{\mathbb{I}(A = a_k)}{\hat{\pi}_0(a_k|X)} \Big(\prod_{j=1}^{k-1} \frac{\hat{f}_j(M_j|X, a_j, \overline{M}_{j-1})}{\hat{f}_j(M_j|X, a_k, \overline{M}_{j-1})} \Big) \big(\hat{\mu}_k^{\rm mle}(X, \overline{M}_k) - \hat{\mu}_{k-1}^{\rm mle}(X, \overline{M}_{k-1}) \big) \\ &+ \hat{\mu}_0^{\rm mle}(X)\Big],\end{aligned} \tag{12}$$

where $\hat{\mu}_k^{\rm mle}(X, \overline{M}_k)$ is iteratively constructed as

$$\hat{\mu}_k^{\rm mle}(X, \overline{M}_k) = \int \hat{\mu}_{k+1}^{\rm mle}(X, \overline{M}_k, m_{k+1}) \hat{f}_{k+1}(m_{k+1}|X, a_{k+1}, \overline{M}_k) d\nu(m_{k+1}), \quad k = K-1, \ldots 0. \tag{13}$$

When M_{k+1} involves continuous components, the above integral can be evaluated via Monte Carlo simulation.

When some of the mediators are continuous/multivariate, however, it can be difficult to estimate the conditional distributions $p(m_k|x, a, \overline{m}_{k-1})$. 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 $\mu_k(x, \overline{m}_k)$ using the regression-imputation approach. Specifically, suppose we have estimated $2 \cdot (K+1)$ nuisance functions: $\hat{\pi}_0(a|x)$ for $p(a|x)$, $\hat{\pi}_k(a|x, \overline{m}_k)$ for $p(a|x, \overline{m}_k)$ ($k \in [K]$), and $\hat{\mu}_k(x, \overline{m}_k)$ for $\mu_k(x, \overline{m}_k)$ ($k \in \{0, 1, \dots K\}$), where for $k < K$, $\mu_k(x, \overline{m}_k)$ is estimated iteratively by fitting a model for the conditional mean of $\hat{\mu}_{k+1}(X, \overline{M}_{k+1})$ given $(X, A, \overline{M}_k)$ and then setting $A = a_{k+1}$ for all units. The GMF can then be estimated as

$$
\begin{aligned}
\hat{\psi}_{\overline{a}}^{\mathrm{eif}_2} =& \mathbb{P}_n \Big[\frac{\mathbb{I}(A = a_{K+1})}{\hat{\pi}_0(a_1|X)} \Big(\prod_{j=1}^{K} \frac{\hat{\pi}_j(a_j|X, \overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X, \overline{M}_j)} \Big) \big(Y - \hat{\mu}_K(X, \overline{M}_K)\big) \\
&+ \sum_{k=1}^{K} \frac{\mathbb{I}(A = a_k)}{\hat{\pi}_0(a_1|X)} \Big(\prod_{j=1}^{k-1} \frac{\hat{\pi}_j(a_j|X, \overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X, \overline{M}_j)} \Big) \big(\hat{\mu}_k(X, \overline{M}_k) - \hat{\mu}_{k-1}(X, \overline{M}_{k-1})\big) \\
&+ \hat{\mu}_0(X) \Big].
\end{aligned} \tag{14}
$$

The multiple robustness and semiparametric efficiency of $\hat{\psi}_{\overline{a}}^{\mathrm{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\mathrm{eif}_2}$ are given below.

Theorem 3. *Let $\eta_1 = \{\pi_0, f_1, \dots f_K, \mu_K\}$ denote the $K+2$ nuisance functions involved in $\hat{\psi}_{\overline{a}}^{\mathrm{eif}_1}$, and $\hat{\eta}_1 = \{\hat{\pi}_0, \hat{f}_1, \dots \hat{f}_K, \hat{\mu}_K\}$ their estimates. Suppose that assumptions 1*-3* and suitable regularity conditions for estimating equations (e.g., Newey and McFadden 1994, p. 2148) hold.*

1. *$\hat{\psi}_{\overline{a}}^{\mathrm{eif}_1}$ is consistent and asymptotically normal (CAN) if any $K+1$ of the $K+2$ nuisance functions in η_1 are correctly specified and their parameter estimates are $\sqrt{n}$-consistent; and it is semiparametric efficient if all of the $K+2$ nuisance functions in η_1 are correctly specified and their parameter estimates are $\sqrt{n}$-consistent.*

2. *$\hat{\psi}_{\overline{a}}^{\mathrm{eif}_2}$ is CAN if the first k treatment models $\pi_0, \dots \pi_{k-1}$ and the last $K+1-k$ outcome models $\mu_k, \dots \mu_K$ are correctly specified and their parameter estimates are $\sqrt{n}$-consistent, where $k \in \{0, \dots K+1\}$; and it is semiparametric efficient if all of the treatment and outcome models are correctly specified and their parameter estimates are $\sqrt{n}$-consistent.*

Both $\hat{\psi}_{\overline{a}}^{\mathrm{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\mathrm{eif}_2}$ are $K+2$-robust in the sense that they are CAN if any of $K+2$ sets of nuisance functions are correctly specified and their parameter estimates are $\sqrt{n}$-consistent. Several special cases are worth mentioning. First, in the degenerate case where $K = 0$, it is clear that both $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ reduce to the standard doubly robust estimator for $\mathbb{E}[Y(a)]$ (Scharfstein *et al.* 1999). Second, when $K = 1$, $\hat{\psi}_{10}^{\text{eif}_1}$ coincides with Tchetgen Tchetgen and Shpitser's (2012) triply robust estimator for $\mathbb{E}[Y\big(0, M(1)\big)]$. Finally, when $K = 2$, $\hat{\psi}_{010}^{\text{eif}_1}$ is identical to Miles *et al.*'s (2020) estimator for ψ_{010}. For this case, however, Miles *et al.* provided a slightly weaker condition than that implied by Theorem 3 for $\hat{\psi}_{010}^{\text{eif}_1}$ to be CAN. Specifically, they showed that $\hat{\psi}_{010}^{\text{eif}_1}$ remains CAN even if both f_1 and μ_2 are misspecified. In Section 4.3, we show that such relaxations of conditions are possible for several particular types of PSEs, including the natural path-specific effect (nPSE), of which $\psi_{010} - \psi_{000}$ is a special case. The $K + 2$-robustness of $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$, interestingly, echoes the multiple robustness of the Bang-Robins (2005) estimator for estimating 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 $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is $K + 2$-robust, let us consider cases in which only one nuisance function in η_1 is misspecified. When only π_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 $\{f_1, \dots f_K, \mu_K\}$ are all correctly specified. When only μ_K is misspecified, all terms involving $\hat{\mu}_K(X, \overline{M}_K)$ and $\hat{\mu}_k^{\text{mle}}(X, \overline{M}_k)$ $(k = 0, 1, \dots K - 1)$ inside $\mathbb{P}_n[\cdot]$ will have a zero mean (asymptotically), leaving only a weighted average of Y (i.e., the weighting estimator (7)), which is consistent because the corresponding nuisance functions $\{\pi_0, f_1, \dots f_K\}$ 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 $\hat{\mu}_k^{\text{mle}}(X, \overline{M}_k)$ $(k < k')$ inside $\mathbb{P}_n[\cdot]$ will have a zero mean (asymptotically), leaving only a weighted average of $\hat{\mu}_{k'}^{\text{mle}}(X, \overline{M}_{k'})$. The latter is one of the "hybrid" estimators mentioned in the previous section, and it is consistent in this case because its corresponding nuisance functions $\{\pi_0, f_1, \dots f_{k'-1}, f_{k'+1}, \dots f_K, \mu_K\}$ are all correctly specified.

The $K + 2$-robustness of $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ is due to a similar logic to that of $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$. Yet, different from $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$, $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ involves estimating $2 \cdot (K + 1)$ nuisance functions, $K + 1$ for the conditional probabilities of treatment and $K + 1$ for the conditional means of observed/imputed outcomes. Also, unlike $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$, the treatment models involved in $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ are not variationally independent; neither are the outcome models. For example, when $M_K \perp\!\!\!\perp A|X, \overline{M}_{K-1}$, $\pi_K(A|X, \overline{M}_K)$ should be identical to $\pi_{K-1}(A|X, \overline{M}_{K-1})$; similarly, when $M_K \perp\!\!\!\perp Y|X, A, \overline{M}_{K-1}$, $\mu_K(X, \overline{M}_K)$ should be identical to $\mu_{K-1}(X, \overline{M}_{K-1})$. Thus, in practice, both the treatment models and the 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 $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ 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 $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$. 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 if the nuisance functions associated with any of the $K+2$ conditions in Theorem 3 are consistently estimated. For $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$, an added advantage of employing data-adaptive methods to estimate the nuisance functions is that, by exploring a larger space within $\mathcal{P}_{\text{np}}$, the risk of model incompatibility is reduced.

When highly-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 (Zheng and van der Laan 2011; Chernozhukov *et al.* 2018; Newey and Robins 2018). 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 fold S_j ("estimation sample"), estimate the target parameter with nuisance functions learned from the remainder of the sample ($S \backslash S_j$; "training sample"); (c) average these fold-specific estimates to form a final estimate of the target parameter. In related work, Newey and Robins (2018) suggest the use of separate training samples to estimate different nuisance functions in step (b), which can lead to faster convergence of higher-order terms.

When cross-fitting is used, $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ 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 $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ 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 $\eta_1 = \{\pi_0, f_1, \dots f_K, \mu_K\}$ denote the $K+2$ nuisance functions involved in $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$, and $\hat{\eta}_1 = \{\hat{\pi}_0, \hat{f}_1, \dots \hat{f}_K, \hat{\mu}_K\}$ their estimates. Let $R_n(\cdot)$ denote a mapping from a nuisance function estimator to its $L_2(P)$ convergence rate where P denotes the true distribution of $O = (X, A, \overline{M}_K, Y)$. Suppose that all assumptions required for Theorem 3 hold. In addition, assume that $\mu_K(X, \overline{M}_K)$ is bounded with probability one. Then, when the nuisance functions are estimated via data-adaptive methods and cross-fitting,*

1. *$\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is consistent if $\sum_{u,v \in \hat{\eta}_1; u \neq v} R_n(u) R_n(v) = o(1)$, and it is semiparametric efficient if $\sum_{u,v \in \hat{\eta}_1; u \neq v} R_n(u) R_n(v) = o(n^{-1/2})$;*

2. *$\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ is consistent if $\sum_{j=0}^{K} R_n(\hat{\pi}_j) R_n(\hat{\mu}_j) = o(1)$, and it is semiparametric efficient if $\sum_{j=0}^{K} R_n(\hat{\pi}_j) R_n(\hat{\mu}_j) = o(n^{-1/2})$.*

The first part of Theorem 4 implies that $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is consistent if any $K+1$ of the $K+2$ nuisance functions are consistently estimated, echoing Theorem 3. Moreover, it suggests that $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is semiparametric efficient if for any two nuisance functions in η_1, the product of their convergence rates is $o(n^{-1/2})$. Thus $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is semiparametric efficient if all of the $K+2$ nuisance functions are consistently estimated and converge at faster than $n^{-1/4}$ rates, but it will also attain the semiparametric efficiency bound under alternative conditions — for example, when estimates of the treatment and mediator models ($\hat{\pi}_0$, $\hat{f}_1, \dots \hat{f}_K$) all converge at a rate of $n^{-1/3}$ and estimates of the outcome model $\hat{\mu}_K$ converge at a rate of $n^{-1/5}$.

The second part of Theorem 4 appears to suggest that the consistency and asymptotic efficiency of $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ require weaker conditions, which involve the sum of only $K+1$, rather than $\binom{K+2}{2}$, product terms. However, because the outcome models are estimated iteratively, the convergence rate of $\hat{\mu}_k$ will in general depend on the convergence rates of $\{\hat{\mu}_{k+1}, \dots \hat{\mu}_K\}$. For example, if $R_n(\hat{\mu}_{k+1}) = O(n^{\delta})$, $R_n(\hat{\mu}_k)$ is unlikely to be faster than $O(n^{\delta})$. Thus, the condition $\sum_{j=0}^{K} R_n(\hat{\pi}_j) R_n(\hat{\mu}_j) = o(1)$ suggests that the first k treatment models $\pi_0, \dots \pi_{k-1}$ and the last $K+1-k$ outcome models $\mu_k, \dots \mu_K$ need to be correctly specified and consistently estimated, where $k \in \{0, \dots K+1\}$. Nonetheless, $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ will be semiparametric efficient under relatively weak conditions, for example, when estimates of the treatment models all converge at a rate of $n^{-1/3}$ and estimates of the outcome models all converge at a rate of $n^{-1/5}$.

3.3 Stable Estimation in Finite Samples

Since both of the multiply robust estimators described above involve inverse probability weights, estimates of the corresponding PSEs may be unstable in situations where the positivity assumption is nearly violated (Kang and Schafer 2007; Petersen *et al.* 2012). A variety of methods have been proposed to improve the finite-sample performance of doubly robust and multiply robust estimators under similar settings (e.g., Robins *et al.* 2007; Tchetgen Tchetgen and Shpitser 2012). Among them, a popular 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 imputation or "substitution" estimator that typically resides in the parameter space of the estimand. Below, we briefly describe how this approach can be adapted to $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$.

Let us start with $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$, which can be written as

$$\begin{aligned}\hat{\psi}_{\overline{a}}^{\text{eif}_2} =& \mathbb{P}_n\big[\hat{w}_K(X, A, \overline{M}_K)\big(Y - \hat{\mu}_K(X, \overline{M}_K)\big)\\ &+ \sum_{k=1}^{K} \hat{w}_{k-1}(X, A, \overline{M}_{k-1})\big(\hat{\mu}_k(X, \overline{M}_k) - \hat{\mu}_{k-1}(X, \overline{M}_{k-1})\big)\\ &+ \hat{\mu}_0(X)\big],\end{aligned} \tag{15}$$

where $\hat{w}_k(A, X, \overline{M}_k)$ $(0 \leq k \leq K)$ are estimates of the corresponding inverse probability weights as displayed in equation (14). Note that the nuisance functions $\hat{\mu}_k(X, \overline{M}_k)$ $(0 \leq k \leq K)$ here are all estimated via the regression-imputation approach. When the corresponding outcome models are fitted via generalized linear models (GLM) with canonical links, one can either (a) fit weighted GLMs (with an intercept term) for $\hat{\mu}_k(X, \overline{M}_k)$ using $\hat{w}_k(A, X, \overline{M}_k)$ as weights, respectively, 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 $\psi_{\overline{a}}$ if it 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 and Rubin 2006; Zheng and 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 $k = K, \dots 0$

   (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 regression-imputation 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), obtain an updated estimate $\hat{\mu}_k^{\text{tmle}}(X, \overline{M}_k) = g^{-1}\big(g(\hat{\mu}_k(X, \overline{M}_k)) + \hat{\beta}_k \hat{w}_k(A, X, \overline{M}_k)\big)$, 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 $\hat{\psi}_{\overline{a}}^{\text{tmle}} = \mathbb{P}_n[\hat{\mu}_0^{\text{tmle}}(X)]$.

As with the GLM-based adjustments, the TMLE approach also yields a substitution estimator that resides in the parameter space of $\psi_{\overline{a}}$ if it equals the range of the model specified for $\mu_0(x)$. It should be noted, however, 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 training sample $(S \backslash S_j)$ and steps 1(b) implemented in the estimation sample S_j. The method of TMLE can also be used to adjust $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$, 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 $\psi_{\overline{a}}$ for the unconstrained case where $a_1, \dots a_{K+1}$ can each take 0 or 1. In many applications, the researcher may be interested in particular causal estimands such as the natural direct effect (NDE), the natural indirect effect (NIE), and natural path-specific effects (nPSE; Daniel *et al.* 2015). Below, we discuss how the multiply robust semiparametric estimators of $\psi_{\overline{a}}$ apply to these estimands. In addition, we discuss a set of cumulative path-specific effects (cPSEs) that together compose the ATE, and how they relate to noncausal decompositions of between-group disparities that are commonly used in the social sciences. For illustrative purposes, we focus on the estimator $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$, although similar results hold for $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$.

4.1 Natural Direct Effect (NDE)

The NDE measures the effect of switching treatment status from 0 to 1 in the hypothetical scenario where the mediators $(M_1, \dots M_K)$ were all set to values they would have "naturally" taken for each unit under treatment status $A = 0$. It is thus given by $\psi_{\overline{0}_K,1} - \psi_{\overline{0}_{K+1}}$, for which a semiparametric efficient estimator can be constructed as

$$\widehat{\text{NDE}}^{\text{eif}_2} = \hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2} - \hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}. \tag{16}$$

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

By setting $a_1 = \dots a_{K+1} = 0$ in equation (14), we have

$$\hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2} = \mathbb{P}_n\big[\frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0|X)}\big(Y - \hat{\mu}_0(X)\big) + \hat{\mu}_0(X)\big], \tag{17}$$

where $\mu_0(X) = \mathbb{E}[Y|X, A = 0]$. Not surprisingly, $\hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}$ 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 $a_1 = \dots a_K = 0$ and $a_{K+1} = 1$ in equation (14), we have

$$\hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2} = \mathbb{P}_n\big[\frac{\mathbb{I}(A=1)}{\hat{\pi}_0(0|X)}\frac{\hat{\pi}_K(0|X,\overline{M}_K)}{\hat{\pi}_K(1|X,\overline{M}_K)}\big(Y - \hat{\mu}_K(X,\overline{M}_K)\big) + \frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0|X)}\big(\hat{\mu}_K(X,\overline{M}_K) - \hat{\mu}_{0,K+1}(X)\big) + \hat{\mu}_{0,K+1}(X)\big].$$

In contrast to the general case where $\overline{a}_K$ is unconstrained, $\hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2}$ involves estimating only four nuisance functions: $\pi_0(a|x)$, $\pi_K(a|x,\overline{m}_K)$, $\mu_{0,K+1}(x)$, and $\mu_K(x,\overline{m}_K)$, where $\mu_K(x,\overline{m}_K) = \mathbb{E}[Y|x, A = 1, \overline{m}_K]$ and $\mu_{0,K+1}(x) = \mathbb{E}[\mu_K(X,\overline{M}_K)|x, A = 0]$. Hence $\mu_{0,K+1}(x)$ 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 $\hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2}$ is triply robust in that it is consistent if any of the following three conditions hold: (a) $\hat{\pi}_0$ and $\hat{\pi}_K$ are consistent; (b) $\hat{\pi}_0$ and $\hat{\mu}_K$ are consistent; and (c) $\hat{\mu}_{0,K+1}$ and $\hat{\mu}_K$ are consistent. In the meantime, we know that $\hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}$ is consistent if either $\hat{\pi}_0$ or $\hat{\mu}_0$ is consistent. By taking the intersection of the multiple robustness conditions for $\hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2}$ and $\hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}$, we deduce that $\widehat{\text{NDE}}^{\text{eif}_2}$ 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, $\widehat{\text{NDE}}^{\text{eif}_2}$ is CAN if any of the following three sets of nuisance functions are correctly specified and their parameter estimates $\sqrt{n}$-consistent: $\{\pi_0, \pi_K\}$, $\{\pi_0, \mu_K\}$, $\{\mu_0, \mu_{0,K+1}, \mu_K\}$. $\widehat{\text{NDE}}^{\text{eif}_2}$ 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, $\widehat{\text{NDE}}^{\text{eif}_2}$ is semiparametric efficient if $R_n(\hat{\pi}_0)R_n(\hat{\mu}_{0,K+1}) + R_n(\hat{\pi}_K)R_n(\hat{\mu}_K) + R_n(\hat{\pi}_0)R_n(\hat{\mu}_0) = o(n^{-1/2})$.*

4.2 Natural Indirect Effect (NIE) for M_1

In section 2.1, we noted that $\psi_{100} - \psi_{000}$ corresponds to the NIE for the first mediator M_1. This correspondence naturally extends to the case of K mediators, where the NIE for M_1 is given by

$$\text{NIE}_{M_1} = \psi_{1,\underline{0}_2} - \psi_{\overline{0}_{K+1}},$$

where $\underline{0}_2 = (0, \ldots 0)$ is a vector of length K representing the fact that $a_2 = \ldots = a_{K+1} = 0$. Thus, a semiparametric efficient estimator of NIE_{M_1} can be constructed as

$$\widehat{\text{NIE}}_{M_1}^{\text{eif}_2} = \hat{\psi}_{1,\underline{0}_2}^{\text{eif}_2} - \hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}.$$

If, instead, we use $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ in the above equation, we obtain Tchetgen Tchetgen and Shpitser's (2012) estimator of NIE_{M_1}.

As shown previously, $\hat{\psi}_{\overline{0}_{K+1}}^{\text{eif}_2}$ is given by the doubly robust estimator (17). Similarly, by setting $a_1 = 1$ and $a_2 = \ldots a_{K+1} = 0$ in equation (14), we obtain

$$\hat{\psi}_{1,\underline{0}_2}^{\text{eif}_2} = \mathbb{P}_n\Big[\frac{\mathbb{I}(A=0)}{\hat{\pi}_0(1|X)}\frac{\hat{\pi}_1(1|X,M_1)}{\hat{\pi}_1(0|X,M_1)}\big(Y - \hat{\mu}_1(X,M_1)\big) + \frac{\mathbb{I}(A=1)}{\hat{\pi}_0(1|X)}\big(\hat{\mu}_1(X,M_1) - \hat{\mu}_{0,1}(X)\big) + \hat{\mu}_{0,1}(X)\Big].$$

Like $\hat{\psi}_{\overline{0}_K,1}^{\text{eif}_2}$, $\hat{\psi}_{1,\underline{0}_2}^{\text{eif}_2}$ 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 $\hat{\psi}_{1,\underline{0}_2}^{\text{eif}_2}$ is triply robust in the sense that it is consistent if any of the following three conditions hold: (a) $\hat{\pi}_0$ and $\hat{\pi}_1$ are consistent; (b) $\hat{\pi}_0$ and $\hat{\mu}_1$ are consistent; and (c) $\hat{\mu}_{0,1}$ and $\hat{\mu}_1$ are consistent. By taking the intersection of the multiple robustness conditions for $\hat{\psi}_{1,\underline{0}_2}^{\text{eif}_2}$ and

$\hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}}$, we deduce that $\widehat{\text{NIE}}^{\text{eif}_2}_{M_1}$ 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, $\widehat{\text{NIE}}^{\text{eif}_2}_{M_1}$ is CAN if any of the following three sets of nuisance functions are correctly specified and their parameter estimates $\sqrt{n}$-consistent: $\{\pi_0, \pi_1\}$, $\{\pi_0, \mu_1\}$, $\{\mu_0, \mu_{0,1}, \mu_1\}$. $\widehat{\text{NIE}}^{\text{eif}_2}_{M_1}$ 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, $\widehat{\text{NIE}}^{\text{eif}_2}_{M_1}$ is semiparametric efficient if $R_n(\hat{\pi}_0)R_n(\hat{\mu}_{0,1}) + R_n(\hat{\pi}_1)R_n(\hat{\mu}_1) + R_n(\hat{\pi}_0)R_n(\hat{\mu}_0) = o(n^{-1/2})$.*

4.3 Natural Path-Specific Effects (nPSE) 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

$$\text{nPSE}_{M_k} = \psi_{\overline{0}_{k-1},1,\underline{0}_{k+1}} - \psi_{\overline{0}_{K+1}}.$$

It can be interpreted as the effect of activating the path $A \to M_k \rightsquigarrow Y$ while all other causal paths are "switched off." A semiparametric efficient estimator of nPSE_{M_k} can be constructed as

$$\widehat{\text{nPSE}}^{\text{eif}_2}_{M_k} = \hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},1,\underline{0}_{k+1}} - \hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}}.$$

If, instead, we use $\hat{\psi}^{\text{eif}_1}_{\overline{a}}$ in the above equation, the resulting estimator $\widehat{\text{nPSE}}^{\text{eif}_1}_{M_k}$ can be seen as an extension of Miles *et al.*'s (2020) estimator of ψ_{010}.

Again, $\hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}}$ is given by the doubly robust estimator (17). By setting $a_1 = \dots a_{k-1} = a_{k+1} = \dots a_{K+1} = 0$ and $a_k = 1$ in equation (14), we obtain

$$\begin{aligned}\hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},1,\underline{0}_{k+1}} =& \mathbb{P}_n\Big[\frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0|X)}\frac{\hat{\pi}_{k-1}(0|X,\overline{M}_{k-1})}{\hat{\pi}_{k-1}(1|X,\overline{M}_{k-1})}\frac{\hat{\pi}_k(1|X,\overline{M}_k)}{\hat{\pi}_k(0|X,\overline{M}_k)}\big(Y - \hat{\mu}_k(X,\overline{M}_k)\big). \\ &+ \frac{\mathbb{I}(A=1)}{\hat{\pi}_0(0|X)}\frac{\hat{\pi}_{k-1}(0|X,\overline{M}_{k-1})}{\hat{\pi}_{k-1}(1|X,\overline{M}_{k-1})}\big(\hat{\mu}_k(X,\overline{M}_k) - \hat{\mu}_{k-1}(X,\overline{M}_{k-1})\big) \\ &+ \frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0|X)}\big(\hat{\mu}_{k-1}(X,\overline{M}_{k-1}) - \hat{\mu}_{0,k}(X)\big) + \hat{\mu}_{0,k}(X)\Big].\end{aligned}$$

We can see that $\hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},1,\underline{0}_{k+1}}$ involves estimating six nuisance functions: $\pi_0(a|x)$, $\pi_{k-1}(a|x,\overline{m}_{k-1})$, $\pi_k(a|x,\overline{m}_k)$, $\mu_0(x)$, $\mu_{k-1}(x,\overline{m}_{k-1})$, and $\mu_k(x,\overline{m}_k)$, where $\mu_k(X,\overline{M}_k) = \mathbb{E}[Y|X,A=0,\overline{M}_k]$,

$\mu_{k-1}(X, \overline{M}_{k-1}) = \mathbb{E}[\mu_k(X, \overline{M}_k)|X, A = 1, \overline{M}_{k-1}]$, and $\mu_{0,k}(X) = \mathbb{E}[\mu_{k-1}(X, \overline{M}_{k-1})|X, A = 0]$. Hence $\mu_{k-1}(x)$ can be estimated by fitting a model for the conditional mean of $\hat{\mu}_k(X, \overline{M}_k)$ given $(X, A, \overline{M}_{k-1})$ and then setting $A = 1$ for all units, and $\mu_{0,k}(x)$ can be estimated by fitting a model for the conditional mean of $\hat{\mu}_{k-1}(X, \overline{M}_{k-1})$ given (X, A) and then setting $A = 0$ for all units. It follows from Theorem 3 that $\hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},1,\underline{0}_{k+1}}$ is quadruply robust in that it is consistent if any of the following four conditions hold: (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) $\hat{\pi}_0$, $\hat{\mu}_{k-1}$, and $\hat{\mu}_k$ are consistent; and (d) $\hat{\mu}_{0,k}$, $\hat{\mu}_{k-1}$, and $\hat{\mu}_k$ are consistent. By taking the intersection of the multiple robustness conditions for $\hat{\psi}^{\text{eif}_2}_{1,\underline{0}_2}$ and $\hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}}$, we deduce that $\widehat{\text{nPSE}}^{\text{eif}_2}_{M_k}$ 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,* $\widehat{\text{nPSE}}^{\text{eif}_2}_{M_k}$ *is CAN if any of the following four sets of nuisance functions are correctly specified and their parameter estimates* $\sqrt{n}$*-consistent:* $\{\pi_0, \pi_{k-1}, \pi_k\}$, $\{\pi_0, \pi_{k-1}, \mu_k\}$, $\{\pi_0, \mu_{k-1}, \mu_k\}$, $\{\mu_0, \mu_{0,k}, \mu_{k-1}, \mu_k\}$. $\widehat{\text{nPSE}}^{\text{eif}_2}_{M_k}$ *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,* $\widehat{\text{nPSE}}^{\text{eif}_2}_{M_k}$ *is semiparametric efficient if* $R_n(\hat{\pi}_0)R_n(\hat{\mu}_{0,k}) + R_n(\hat{\pi}_{k-1})R_n(\hat{\mu}_{k-1}) + R_n(\hat{\pi}_k)R_n(\hat{\mu}_k) + R_n(\hat{\pi}_0)R_n(\hat{\mu}_0) = o(n^{-1/2})$.

4.4 Cumulative Path-Specific Effects (cPSE) for M_k $(k \geq 2)$

The NDE, NIE, and nPSE 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 $\psi_{\overline{0}_K,1} - \psi_{\overline{0}_{K+1}}$equals the NDE, and the component $\psi_{\overline{0}_{k-1},\underline{1}_k} - \psi_{\overline{0}_k,\underline{1}_{k+1}}$ gauges the additional contribution of the causal path $A \to M_k \rightsquigarrow Y$ after the causal paths going through the mediators $M_{k+1}, \ldots M_K$, as well as the path $A \to Y$, are switched on. Such a decomposition can be useful in applications where the investigator aims to partition the ATE into its path-specific components.

For the decomposition given by equation (3), we define the cumulative path-specific effect (cPSE) for mediator M_k ($k \geq 2$) as

$$\text{cPSE}_{M_k} = \psi_{\overline{0}_{k-1},\underline{1}_k} - \psi_{\overline{0}_k,\underline{1}_{k+1}},$$

for which a semiparametric efficient estimator can be constructed as

$$\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k} = \hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},\underline{1}_k} - \hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}}.$$

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

$$\hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}} = \mathbb{P}_n\Big[\frac{\mathbb{I}(A=1)}{\hat{\pi}_0(0|X)}\frac{\hat{\pi}_k(0|X,\overline{M}_k)}{\hat{\pi}_k(1|X,\overline{M}_k)}\big(Y - \hat{\mu}_k(X,\overline{M}_k)\big) + \frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0|X)}\big(\hat{\mu}_k(X,\overline{M}_k) - \hat{\mu}_{0,k}(X)\big) + \hat{\mu}_{0,k}(X)\Big], \tag{18}$$

where $\mu_k(X,\overline{M}_k) = \mathbb{E}[Y|X,A=1,\overline{M}_k]$ and $\mu_{0,k}(X) = \mathbb{E}[\mu_k(X,\overline{M}_k)|X,A=0]$. Note that the definition of $\mu_{0,k}(X)$ here differs from that in the previous section. It follows from Theorem 3 that $\hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}}$ is triply robust in that it is consistent if any of the following three conditions hold: (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 k with $k-1$ in equation (18), we can construct a similar expression for $\hat{\psi}^{\text{eif}_2}_{\overline{0}_{k-1},\underline{1}_k}$, which is also triply robust in that it is consistent if any of the following three conditions hold: (a) $\hat{\pi}_0$ and $\hat{\pi}_{k-1}$ are consistent; (b) $\hat{\pi}_0$ and $\hat{\mu}_{k-1}$ are consistent; and (c) $\hat{\mu}_{0,k-1}$ and $\hat{\mu}_{k-1}$ are consistent. As a result, $\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k}$ involves fitting seven working models — for $\pi_0(a|x)$, $\pi_{k-1}(a|x,\overline{m}_{k-1})$, $\pi_k(a|x,\overline{m}_k)$, $\mu_{k-1}(x)$, $\mu_{0,k-1}(x,\overline{m}_{k-1})$, $\mu_k(x)$, and $\mu_{0,k}(x,\overline{m}_k)$. By taking the intersection of the multiple robustness conditions for $\hat{\psi}^{\text{eif}_2}_{\overline{1}_k,\underline{0}_{k+1}}$ and $\hat{\psi}^{\text{eif}_2}_{\overline{1}_{k-1},\underline{0}_k}$, we deduce that $\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k}$ is quintuply robust in that it is consistent when any of five sets of nuisance functions are 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, $\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k}$ is CAN if any of the following five sets of nuisance functions are correctly specified and their parameter estimates $\sqrt{n}$-consistent: $\{\pi_0,\pi_{k-1},\pi_k\}$;$\{\pi_0,\pi_{k-1},\mu_k\}$;$\{\pi_0,\mu_{k-1},\pi_k\}$;$\{\pi_0,\mu_{k-1},\mu_k\}$;$\{\mu_{0,k-1},\mu_{0,k},\mu_{k-1},\mu_k\}$. $\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k}$ 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, $\widehat{\text{cPSE}}^{\text{eif}_2}_{M_k}$ is semiparametric efficient if $R_n(\hat{\pi}_0)R_n(\hat{\mu}_{0,k-1})+$*

$R_n(\hat{\pi}_0)R_n(\hat{\mu}_{0,k}) + R_n(\hat{\pi}_{k-1})R_n(\hat{\mu}_{k-1}) + R_n(\hat{\pi}_k)R_n(\hat{\mu}_k) = o(n^{-1/2}).$

4.5 Decomposition of Between-group Disparities

Equation (3) can also be employed to decompose between-group disparities in a scalar outcome, such as the black-white income gap, into components that are attributable to group differences in various ascriptive and achieved characteristics (Fortin *et al.* 2011). For example, using linear structural equation models, Duncan (1968) decomposed the total black-white income gap into components that reflect black-white differences in family background, academic performance (net of family background), educational attainment (net of family background and academic performance), occupational attainment (net of family background, academic performance, and educational attainment), and a "residual" component that cannot be explained by the above characteristics. Although proposed prior to Blinder (1973) and Oaxaca (1973), Duncan's decomposition can be seen as a generalization of the Blinder-Oaxaca decomposition widely used in labor economics. Algebraically, this decomposition is equivalent to equation (3) except that the left-hand side is now the black-white income gap rather than the average causal effect of a manipulable intervention, and, as a result, there are no pretreatment confounders X. However, this decomposition is different from causal mediation analysis for a randomized trial, in which case pretreatment covariates may still be needed to adjust for potential confounding of the mediator-mediator and mediator-outcome relationships. The components associated with Duncan's decomposition, by contrast, are purely statistical parameters and should not be interpreted causally.

Consequently, in the context of decomposing between-group disparities, the functional $\psi_{\overline{0}_k,\underline{1}_{k+1}}$ can be estimated as

$$\hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}} = \mathbb{P}_n\big[\frac{\mathbb{I}(A=1)}{\hat{\pi}_0(0)}\frac{\hat{\pi}_k(0|\overline{M}_k)}{\hat{\pi}_k(1|\overline{M}_k)}\big(Y - \hat{\mu}_k(\overline{M}_k)\big) + \frac{\mathbb{I}(A=0)}{\hat{\pi}_0(0)}\big(\hat{\mu}_k(\overline{M}_k) - \hat{\mu}_{0,k}\big) + \hat{\mu}_{0,k}\big], \tag{19}$$

where $\pi_0(0) = \Pr[A=0]$, $\mu_k(\overline{M}_k) = \mathbb{E}[Y|A=1,\overline{M}_k]$, and $\mu_{0,k} = \mathbb{E}[\mu_k(\overline{M}_k)|A=0]$. Since $\hat{\pi}_0(0)$ can be estimated by the sample average of $1-A$ and $\mu_{0,k}$ the sample average of $\hat{\mu}_k(\overline{M}_k)$ among units with $A=0$, equation (19) involves estimating only two nuisance functions: $\hat{\pi}_k(a|\overline{m}_k)$ and $\hat{\mu}_k(\overline{m}_k)$. It follows from Theorem 3 that $\hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}}$ is now doubly robust — it is consistent if either $\hat{\pi}_k(a|\overline{m}_k)$ or $\hat{\mu}_k(\overline{m}_k)$ is consistent.

To implement the full decomposition given by equation (3), we need to estimate $\psi_{\overline{0}_k,\underline{1}_{k+1}}$ for each $k \in 0, 1, \ldots K+1$, i.e., estimate the vector-valued parameter $\boldsymbol{\psi}_{\text{decomp}} = (\psi_{\underline{1}_1}, \psi_{0,\underline{1}_2}, \ldots \psi_{\overline{0}_K,1}, \psi_{\overline{0}_{K+1}})$. Since $\psi_{\underline{1}_1}$ and $\psi_{\overline{0}_{K+1}}$ can be estimated by the sample analogs of $\mathbb{E}[Y|A=1]$ and $\mathbb{E}[Y|A=0]$ and $\hat{\psi}^{\text{eif}_2}_{\overline{0}_k,\underline{1}_{k+1}}$ is doubly robust with respect to $\hat{\pi}_k$ and $\hat{\mu}_k$, the semiparametric estimator $\hat{\boldsymbol{\psi}}^{\text{eif}_2}_{\text{decomp}} = (\hat{\psi}^{\text{eif}_2}_{\underline{1}_1}, \hat{\psi}^{\text{eif}_2}_{0,\underline{1}_2}, \ldots \hat{\psi}^{\text{eif}_2}_{\overline{0}_K,1}, \hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}})$ is 2^K-robust: it is consistent if for each $k \in [K]$, either $\hat{\pi}_k$ or $\hat{\mu}_k$ is consistent. Note that in this case, the functions $\mu_k(\overline{M}_k) = \mathbb{E}[Y|A=1, \overline{M}_k]$ are not estimated iteratively, but separately for each k.

Corollary 5. *Define $\hat{\boldsymbol{\psi}}^{\text{eif}_2}_{\text{decomp}} = (\hat{\psi}^{\text{eif}_2}_{\underline{1}_1}, \hat{\psi}^{\text{eif}_2}_{0,\underline{1}_2}, \ldots \hat{\psi}^{\text{eif}_2}_{\overline{0}_K,1}, \hat{\psi}^{\text{eif}_2}_{\overline{0}_{K+1}})$. Suppose $X = \varnothing$, and that all assumptions required for Theorem 4 hold. When the nuisance functions $(\hat{\pi}_1, \ldots \hat{\pi}_K, \hat{\mu}_1, \ldots \hat{\mu}_K)$ are estimated via parametric models, $\hat{\boldsymbol{\psi}}^{\text{eif}_2}_{\text{decomp}}$ is CAN if for each $k \in [K]$, either $\hat{\pi}_k$ or $\hat{\mu}_k$ is correctly specified and their estimates $\sqrt{n}$-consistent. $\hat{\boldsymbol{\psi}}^{\text{eif}_2}_{\text{decomp}}$ is semiparametric efficient if all of the 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, $\hat{\boldsymbol{\psi}}^{\text{eif}_2}_{\text{decomp}}$ is semiparametric efficient if $\sum_{k=1}^{K} R_n(\hat{\pi}_k) R_n(\hat{\mu}_k) = o(n^{-1/2})$.*

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, a pretreatment covariate X, and two causally ordered mediators M_1 and M_2. The data generating process is similar to that used in Miles *et al.* (2020) and is described in greater detail in Supplementary Material E. We generate 1,000 Monte Carlo samples of size 2,000, and, without loss of generality, focus on $\psi_{010} = \mathbb{E}[Y(0, M_1(0), M_2(1, M_1(0)))]$, the same estimand studied in Miles *et al.* (2020).

We compare the performance of eight different estimators. First, we consider the weighting estimator $\hat{\psi}^{\text{w-a}}$, the regression-imputation estimator $\hat{\psi}^{\text{ri}}$, and the hybrid estimators $\hat{\psi}^{\text{ri-w-w}}$ and $\hat{\psi}^{\text{ri-ri-w}}$, where the mediator density ratio involved in $\hat{\psi}^{\text{ri-w-w}}$ is estimated via the corresponding odds ratio of the treatment variable. We then consider four EIF-based estimators $\hat{\psi}^{\text{eif}_2}_{\text{par}}$, $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$, $\hat{\psi}^{\text{eif}_2}_{\text{np}}$, and $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$. For $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ and $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$, the nuisance functions are estimated via GLMs. $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$ differs from $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ 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 $\mathbb{P}_n[\cdot]$ but $\hat{\mu}_0(X)$ have a zero sample mean, yielding a substitution estimator that tends to be more stable in finite samples. Among the two nonparametric estimators, $\hat{\psi}^{\text{eif}_2}_{\text{np}}$ is based on estimating equation (14), and $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$ is based on the method of TMLE. Like $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$, $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$ is a substitution estimator, which may have better finite-sample performance than $\hat{\psi}^{\text{eif}_2}_{\text{np}}$.

For both $\hat{\psi}^{\text{eif}_2}_{\text{np}}$ and $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$, 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 corresponding covariates. Two-fold cross-fitting is used to obtain the final estimates of ψ_{010}. Note that in this simulation, the pretreatment covariate X and the mediators M_1 and M_2 are all univariate, making it relatively easy to specify a super learner that spans a sufficiently large model space. In scenarios where one or more of these variables are high-dimensional, the performance of the nonparametric estimators will likely depend on sample size, the data-generating process, as well as the choices of base learners and feature matrices.

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 multiple robustness of $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ and $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$, we consider five different cases: (a) only π_0, π_1, π_2 are correctly specified; (b) only π_0, π_1, μ_2 are correctly specified; (c) only π_0, μ_1, μ_2 are correctly specified; (d) only μ_0, μ_1, μ_2 are correctly specified; and (e) all of the six nuisance functions are misspecified. In theory, $\hat{\psi}^{\text{w-a}}$ is consistent only in case (a), $\hat{\psi}^{\text{ri-w-w}}$ is consistent only in case (b), $\hat{\psi}^{\text{ri-ri-w}}$ is consistent only in case (c), $\hat{\psi}^{\text{ri}}$ is consistent only in case (d), and $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ and $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$ should be consistent in cases (a)-(d).

Results from the simulation study are shown in Figure 2, where each panel corresponds to an estimator, and the y axis is recentered at the true value of $\psi_{0,1,0}$. The shaded box plots highlight the cases under which a given estimator should be consistent. From the first four panels, we can see that the weighting, regression-imputation, and hybrid estimators all behave as expected. They center around the true value if and only if the requisite nuisance functions are all correctly specified. In addition, we observe that in cases where the corresponding nuisance functions are correctly specified, the weighting estimator $\hat{\psi}^{\text{w-a}}$ exhibits a larger variance than the hybrid estimators $\hat{\psi}^{\text{ri-w-w}}$ and $\hat{\psi}^{\text{ri-w-w}}$ and the regression-imputation estimator $\hat{\psi}^{\text{ri}}$.

The next four panels show the box plots of the EIF-based estimators. As expected, both $\hat{\psi}^{\text{eif}_2}_{\text{par}}$

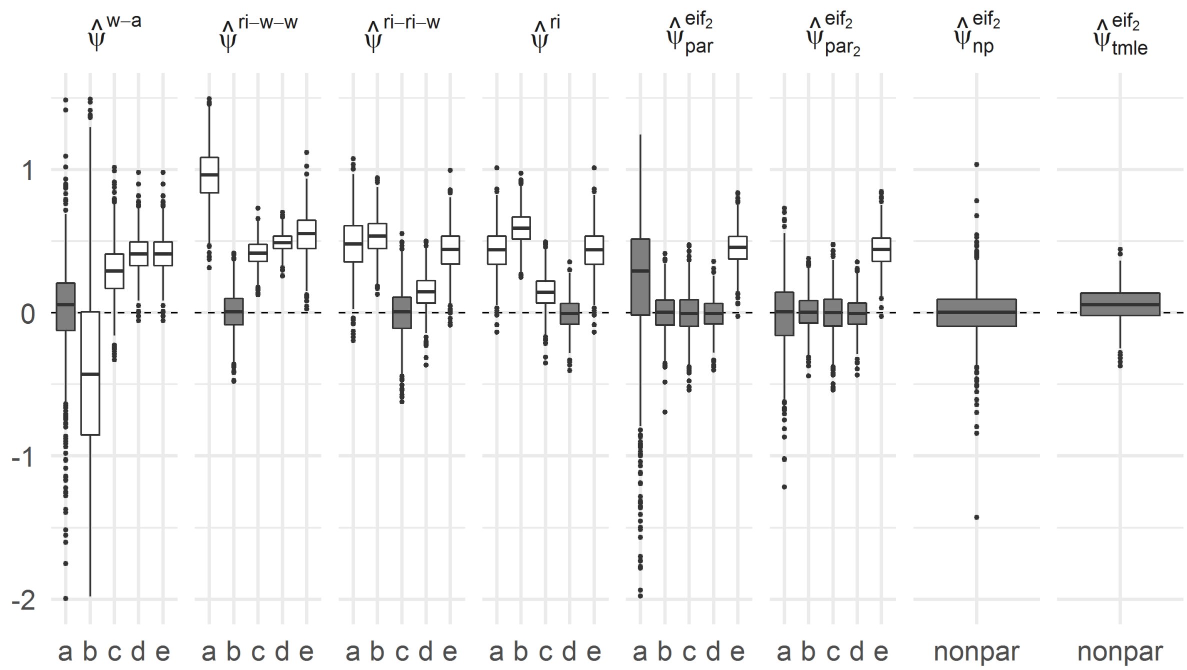

Figure 2: Sampling distributions of eight different estimators for $n = 2,000$. Cases (a)-(d) corresponds to conditions (1)-(4) in Proposition 3, respectively, and case (e) corresponds to a scenario where all nuisance functions are misspecified.

and $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$ are quadruply robust, as their sampling distributions roughly concentrate around the true value in all of the four cases from (a) to (d). In case (a), i.e., when only π_0, π_1, π_2 are correctly specified, $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ exhibits a modest bias and a relatively large variance; by contrast, $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$ is approximately unbiased and much more precise, demonstrating the usefulness of the stabilization technique. Moreover, it is reassuring to see that when all of the six nuisance functions are misspecified (case (e)), $\hat{\psi}^{\text{eif}_2}_{\text{par}}$ and $\hat{\psi}^{\text{eif}_2}_{\text{par}_2}$ do not suffer a larger amount of bias than those of the other parametric estimators. Finally, both of the nonparametric EIF-based estimators perform quite well. The unadjusted estimator $\hat{\psi}^{\text{eif}_2}_{\text{np}}$ is nearly unbiased, but occasionally gives rise to extreme estimates. The TMLE estimator $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$, on the other hand, is slightly biased but more stable than $\hat{\psi}^{\text{eif}_2}_{\text{np}}$. The root mean squared error (RMSE) is 0.19 for $\hat{\psi}^{\text{eif}_2}_{\text{np}}$ and 0.13 for $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$. Their 95% Wald confidence intervals, constructed using the estimated variance $\hat{\mathbb{E}}\big[\big(\hat{\varphi}_{010}(O)\big)^2\big]/n$, have close-to-nominal coverage rates — 93.3% for $\hat{\psi}^{\text{eif}_2}_{\text{np}}$ and 91.5% for $\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$.

6 An Empirical Application

In this section, we illustrate semiparametric estimation of PSEs by analyzing 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.

We use data from $n = 2,969$ individuals in the National Longitudinal Survey of Youth 1997 (NLSY97) who were age 15-17 when they were first interviewed in 1997. The treatment A is a binary indicator for whether the individual attended a two-year or four-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 annual earnings from 2007 to 2010. 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.

To minimize potential bias due to unobserved selection, we include a large number 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 asset, 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, an index of delinquency, whether the respondent had children by age 18, college expectation among the respondent's peers, and a number of school-level characteristics. Some components of X, M_1, and M_2 contain a small fraction of missing values. They are imputed via a random-forest-based

Table 1: Estimates of total and path-specific effects of college attendance on voting.

	Estimating equation ($\hat{\psi}^{\text{eif}_2}_{\text{np}}$)	TMLE ($\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$)
Average total effect	0.153 (0.022)	0.156 (0.023)
Through economic status ($A \to M_1 \rightsquigarrow Y$)	0.009 (0.006)	0.003 (0.006)
Through civic/political interest ($A \to M_2 \to Y$)	0.04 (0.008)	0.047 (0.008)
Direct effect ($A \to Y$)	0.104 (0.021)	0.106 (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.

multiple imputation procedure (with ten 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 $A \to Y$, $A \to M_1 \rightsquigarrow Y$, and $A \to M_2 \to Y$ are identified. For illustrative purposes, we focus on the cumulative PSEs (cPSEs) defined in section 4.4:

$$\text{ATE} = \underbrace{\psi_{001} - \psi_{000}}_{A \to Y} + \underbrace{\psi_{011} - \psi_{001}}_{A \to M_2 \to Y} + \underbrace{\psi_{111} - \psi_{011}}_{A \to M_1 \rightsquigarrow Y}. \tag{20}$$

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 economic status and civic and political interest, respectively. Since both of the mediators are multivariate, it would be difficult to model their conditional distributions directly. We thus estimate the PSEs using the estimator $\hat{\psi}^{\text{eif}_2}_{a_1,a_2,a}$. Each of the nuisance functions is estimated using a super learner composed with Lasso and random forest, where, for computational reasons, the feature matrix 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 ($\hat{\psi}^{\text{eif}_2}_{\text{np}}$), and one based on the method of TMLE ($\hat{\psi}^{\text{eif}_2}_{\text{tmle}}$). Five-fold 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 and path-specific effects. By $\hat{\psi}^{\text{eif}_2}_{\text{np}}$, for example, the estimated total effect of college attendance on voting is 0.153, meaning that on average, college attendance increases the likelihood of voting in 2010 by about 15 percentage points. The estimated PSE via M_2 is 0.04, 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," i.e., 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, we have defined a set of PSEs that can be constructed as contrasts between the expectations of 2^{K+1} potential outcomes, which are identified via what we call the generalized mediation functional (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 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, nPSE, cPSE, as well as noncausal decompositions of between-group disparities. 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," i.e., the degree to which the effect of a treatment operates through multiple mediators sequentially, such as that reflected in the causal path $A \to M_1 \to M_2 \to Y$ (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 and Nelson 2011) or alternative estimands (Lin and VanderWeele 2017). We consider semiparametric estimation and inference for these alternative approaches a promising avenue for future research.

References

Albert, J. M. (2012) Mediation analysis for nonlinear models with confounding. *Epidemiology*, **23**, 879.

Albert, J. M. and Nelson, S. (2011) Generalized causal mediation analysis. *Biometrics*, **67**, 1028–1038.

Alwin, D. F. and Hauser, R. M. (1975) The decomposition of effects in path analysis. *American sociological review*, 37–47.

Avin, C., Shpitser, I. and Pearl, J. (2005) Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 357–363. Morgan Kaufmann Publishers Inc.

Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–973.

Baron, R. M. and 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.

Blinder, A. S. (1973) Wage discrimination: Reduced form and structural estimates. *Journal of Human resources*, 436–455.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**, C1–C68.

Daniel, R., De Stavola, B., Cousens, S. and 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.

Duncan, O. D. (1968) Inheritance of poverty or inheritance of race? *On Understanding Poverty*, 85–110.

Fortin, N., Lemieux, T. and Firpo, S. (2011) Decomposition methods in economics. In *Handbook of labor economics*, vol. 4, 1–102. Elsevier.

Goetgeluk, S., Vansteelandt, S. and Goetghebeur, E. (2009) Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 1049–1066.

Hafeman, D. M. and 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. and 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. and 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. and 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. and Wilson, J. (2003) Why education matters. *Sociology of Education*, 53–70.

Lin, S.-H. and VanderWeele, T. (2017) Interventional approach for path-specific effects. *Journal of Causal Inference*, **5**.

Luedtke, A. R., Sofrygin, O., van der Laan, M. J. and 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. and 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. and Tchetgen Tchetgen, E. J. (2020) On semi-parametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, **107**, 159–172.

Milligan, K., Moretti, E. and 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. and Robins, J. (2017) Multiple robustness in factorized likelihood models. *Biometrika*, **104**, 561–581.

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

Newey, W. K. and 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.

Oaxaca, R. (1973) Male-female wage differentials in urban labor markets. *International Economic Review*, 693–709.

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

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

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

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

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

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

Robins, J. M. and 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.

Rolfe, M. (2012) *Voter Turnout: A Social Theory of Political Participation.* Cambridge University Press.

Rotnitzky, A., Robins, J. and 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.* Wiley and Sons.

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

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

Steen, J., Loeys, T., Moerkerke, B. and 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. and Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, **40**, 1816.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**.

van Der Laan, M. J. and Rubin, D. (2006) Targeted maximum likelihood learning. *The International Journal of Biostatistics*, **2**.

VanderWeele, T. (2015) *Explanation in Causal Inference: Methods for Mediation and Interaction.* New York, NY: Oxford University Press.

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

VanderWeele, T. J. (2009) Concerning the consistency assumption in causal inference. *Epidemiology*, **20**, 880–883.

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

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

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

Wodtke, G. and Zhou, X. (2020) Effect decomposition in the presence of treatment-induced confounding: A regression-with-residuals approach. *Epidemiology.*

Wright, S. (1934) The method of path coefficients. *The annals of mathematical statistics*, **5**, 161–215.

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

Zheng, W. and van der Laan, M. J. (2012) Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics*, **8**.

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

Supplementary Materials

A Hybrid Estimators of $\psi_{a_1,a_2,a}$

For notational brevity, let us use the following shorthands:

$$\lambda_0^j(A|X) \triangleq \frac{\mathbb{I}(A=a_j)}{p(a_j|X)}$$
$$\lambda_1^j(M_1|X) \triangleq \frac{p(M_1|X,a_1)}{p(M_1|X,a_j)}$$
$$\lambda_2^j(M_2|X,M_1) \triangleq \frac{p(M_2|X,a_2,M_1)}{p(M_2|X,a_j,M_1)},$$

In addition, define $\lambda_0(A|X) = \mathbb{I}(A = a)/p(a|X)$, $\lambda_1(M_1|X) = p(M_1|X,a_1)/p(M_1|X,a)$ and $\lambda_2(M_2|X,M_1) = p(M_2|X,a_2,M_1)/p(M_2|X,a,M_1)$. With the above notation, the iterated conditional means $\mu_1(X,M_1)$, $\mu_0(X)$, and $\psi_{a_1,a_2,a}$ can each be written in several different forms:

$$\mu_1(X,M_1) = \begin{cases} \mathbb{E}[\mu_2(X,M_1,M_2)|X,a_2,M_1] \\ \mathbb{E}[\lambda_2(M_2|X,M_1)Y|X,a,M_1] \end{cases}$$

$$\mu_0(X) = \begin{cases} \mathbb{E}[\mu_1(X,M_1)|X,a_1] = \begin{cases} \mathbb{E}\big[\mathbb{E}[\mu_2(X,M_1,M_2)|X,a_2,M_1]|X,a_1\big] \\ \mathbb{E}\big[\mathbb{E}[\lambda_2(M_2|X,M_1)Y|X,a,M_1]|X,a_1\big] \end{cases} \\ \mathbb{E}[\lambda_1^2(M_1|X)\mu_2(X,M_1,M_2)|X,a_2] \\ \mathbb{E}[\lambda_1(M_1|X)\lambda_2(X,M_1,M_2)Y|X,a] \end{cases}$$

$$\psi_{a_1,a_2,a} = \begin{cases} \mathbb{E}[\mu_0(X)] = \begin{cases} \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\mu_2(X,M_1,M_2)|X,a_2,M_1]|X,a_1\big]\Big] & \text{(RI-RI-RI)} \\ \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\lambda_2(M_2|X,M_1)Y|X,a,M_1]|X,a_1\big]\Big] & \text{(W-RI-RI)} \\ \mathbb{E}\Big[\mathbb{E}[\lambda_1^2(M_1|X)\mu_2(X,M_1,M_2)|X,a_2]\Big] & \text{(RI-W-RI)} \\ \mathbb{E}\Big[\mathbb{E}[\lambda_1(M_1|X)\lambda_2(X,M_1,M_2)Y|X,a]\Big] & \text{(W-W-RI)} \end{cases} \\ \mathbb{E}[\lambda_0^1(A|X)\mu_1(X,M_1)] = \begin{cases} \mathbb{E}\big[\lambda_0^1(A|X)\mathbb{E}[\mu_2(X,M_1,M_2)|X,a_2,M_1]\big] & \text{(RI-RI-W)} \\ \mathbb{E}\big[\lambda_0^1(A|X)\mathbb{E}[\lambda_2(M_2|X,M_1)Y|X,a,M_1]\big] & \text{(W-RI-W)} \end{cases} \\ \mathbb{E}[\lambda_0^2(A|X)\lambda_1^2(M_1|X)\mu_2(X,M_1,M_2)] \quad \text{(RI-W-W)} \\ \mathbb{E}[\lambda_0(A|X)\lambda_1(M_1|X)\lambda_2(M_2|X,M_1)Y] \quad \text{(W-W-W)} \end{cases}$$

The first set of equations point to two different ways of estimating $\mu_1(x,m_1)$: (a) fit a model for the conditional mean of $\hat{\mu}_2(X,M_1,M_2)$ given X, A, M_1 and then set $A = a_2$ for all units; (b) fit a model for the conditional mean of $\hat{\lambda}_2(M_2|X,M_1)Y$ given X, A, and M_1 and then set $A = a$ for all units. Similarly, the second set of equations point to four different ways of estimating $\mu_0(x)$, and the last set of equations point to eight different ways of estimating $\psi_{a_1,a_2,a}$. Each of these eight estimators corresponds to a unique combination of regression-imputation and weighting.

B Proof of Theorem 1

Setting $\overline{m}^*_{k-1} = \overline{m}_{k-1}$, assumption 2* implies that

$$\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_k(a_k, \overline{m}_{k-1}) | X, A, \overline{M}_{k-1}. \tag{21}$$

In the meantime, for any $k = 2, \dots K$, we have

$$\begin{aligned}
&\big(M_k(a_k, \overline{m}_{k-1}), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_{k-1}(a_{k-1}, \overline{m}^*_{k-2}) | X, A, \overline{M}_{k-2} \\
\Rightarrow &\big(M_k(a_k, \overline{m}_{k-1}), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_{k-1}(a_{k-1}, \overline{m}^*_{k-2}) | X, A = a_{k-1}, \overline{M}_{k-2} = \overline{m}^*_{k-2} \\
\Rightarrow &\big(M_k(a_k, \overline{m}_{k-1}), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_{k-1} | X, A = a_{k-1}, \overline{M}_{k-2} = \overline{m}^*_{k-2} \\
\Rightarrow &\big(M_k(a_k, \overline{m}_{k-1}), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_{k-1} | X, A, \overline{M}_{k-2}.
\end{aligned} \tag{22}$$

Now suppose that for some $j \in \{1, \dots k-1\}$,

$$\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_k(a_k, \overline{m}_{k-1}) | X, A, \overline{M}_{k-j};$$
$$\big(M_k(a_k, \overline{m}_{k-1}), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_{k-j} | X, A, \overline{M}_{k-j-1}.$$

By the contraction rule of conditional independence, the above relationships imply

$$\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_k(a_k, \overline{m}_{k-1}) | X, A, \overline{M}_{k-j-1}.$$

Hence, by initial relationships (21-22) and mathematical induction, we have

$$\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp M_k(a_k, \overline{m}_{k-1}) | X, A, \quad \forall k \in [K]. \tag{23}$$

In the meantime, because $\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp A | X$, we have (by the contraction rule)

$$\big(M_{k+1}(a_{k+1}, \overline{m}_k), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big) \perp\!\!\!\perp \big(A, M_k(a_k, \overline{m}_{k-1})\big) | X, \quad \forall k \in [K].$$

Thus the components in $\big(A, M_1(a_1), \dots M_K(a_K, \overline{m}_{K-1}), Y(a, \overline{m}_K)\big)$ are mutually independent given X. Therefore,

$$\begin{aligned}
\psi_{\overline{a}} &= \mathbb{E}[Y(a_{K+1}, \overline{M}_K(\overline{a}_K))] \\
&= \int_x \int_{\overline{m}_K} \mathbb{E}[Y(a_{K+1}, \overline{m}_K) | X = x, A = a_{K+1}, M_1(a_1) = m_1, \dots M_K(a_K, \overline{m}_{K-1}) = m_K] \\
&\Big(\prod_{k=1}^K dP_{M_k(a_k, \overline{m}_{k-1}) | X, A, M_1(a_1), \dots M_{k-1}(a_{k-1}, \overline{m}_{k-2})}(m_k | x, a_{K+1}, \overline{m}_{k-1})\Big) dP_X(x)
\end{aligned}$$

$$= \int_x \int_{\overline{m}_K} \mathbb{E}[Y(a_{K+1}, \overline{m}_K)|X = x, A = a_{K+1}]\big(\prod_{k=1}^{K} dP_{M_k(a_k,\overline{m}_{k-1})|X}(m_k|x)\big)dP_X(x)$$

$$= \int_x \int_{\overline{m}_K} \mathbb{E}[Y(a_{K+1}, \overline{m}_K)|x, a_{K+1}, \overline{M}_K = \overline{m}_K]\big(\prod_{k=1}^{K} dP_{M_k(a_k,\overline{m}_{k-1})|X,A,\overline{M}_{k-1}}(m_k|x, a_k, \overline{m}_{k-1})\big)dP_X(x)$$

$$= \int_x \int_{\overline{m}_K} \mathbb{E}[Y|x, a_{K+1}, \overline{m}_K]\big(\prod_{k=1}^{K} dP_{M_k|X,A,\overline{M}_{k-1}}(m_k|x, a_k, \overline{m}_{k-1})\big)dP_X(x).$$

C Proof of Theorem 2

To show that equation (11) is the EIF of $\psi_{\overline{a}}$ in $\mathcal{P}_{\text{np}}$, it suffices to show

$$\left.\frac{\partial \psi_{\overline{a}}(t)}{\partial t}\right|_{t=0} = \mathbb{E}[\varphi_{\overline{a}}(O)S_0(O)], \tag{24}$$

where $S_0(O)$ is the score function for any one-dimensional submodel $P_t(O)$ evaluated at $t = 0$. We first note that $S_t(O)$ can be written as $S_t(O) = S_t(X) + S_t(A|X) + \sum_{k=1}^K S_t(M_k|X, A, \overline{M}_{k-1}) + S_t(Y|X, A, \overline{M}_K)$, where $S_t(u|v) = \partial \log p_t(u|v)/\partial t$ and $p_t(u|v)$ is the conditional probability density/mass function of U given V. Using equation (2) and the product rule, the left hand side of equation (24) can be written as

$$\left.\frac{\partial \psi_{\overline{a}}(t)}{\partial t}\right|_{t=0} = \left.\frac{\partial \iiint y dP_t(y|x, a_{K+1}, \overline{m}_K)\big[\prod_{k=1}^K dP_t(m_k|x, a_k, \overline{m}_{k-1})\big]dP_t(x)}{\partial t}\right|_{t=0}$$

$$= \underbrace{\iiint yS_0(x)dP_0(y|x, a_{K+1}, \overline{m}_K)\big[\prod_{k=1}^{K} dP_0(m_k|x, a_k, \overline{m}_{k-1})\big]dP_0(x)}_{=:\phi_0}$$

$$+ \sum_{k=1}^{K} \underbrace{\iiint yS_0(m_k|x, a_k, \overline{m}_{k-1})dP_0(y|x, a_{K+1}, \overline{m}_K)\big[\prod_{k=1}^{K} dP_0(m_k|x, a_k, \overline{m}_{k-1})\big]dP_0(x)}_{=:\phi_k}$$

$$+ \underbrace{\iiint yS_0(y|x, a_{K+1}, \overline{m}_K)dP_0(y|x, a_{K+1}, \overline{m}_K)\big[\prod_{k=1}^{K} dP_0(m_k|x, a_k, \overline{m}_{k-1})\big]dP_0(x)}_{=:\phi_{K+1}}$$

$$= \sum_{k=0}^{K+1} \phi_k$$

where the second equality follows from the fact that $\partial dP_t(u|v)/\partial t = S_t(u|v)dP_t(u|v)$. Below, we verify that $\phi_k = \mathbb{E}[\varphi_k(O)S_0(O)]$ for all $k \in \{0, \dots K + 1\}$, where $\varphi_k(O)$ is defined in Theorem 2. First,

$$\mathbb{E}[\varphi_0(O)S_0(O)]$$

$$
\begin{aligned}
&=\mathbb{E}[\big(\mu_0(X)-\psi_{\overline{a}}\big)S_0(O)]\\
&=\mathbb{E}[\mu_0(X)S_0(O)]\\
&=\mathbb{E}\big[\mu_0(X)\big(S_0(X)+S_0(A|X)+\sum_{k=1}^{K}S_0(M_k|X,A,\overline{M}_{k-1})+S_0(Y|X,A,\overline{M}_K)\big)\big]\\
&=\mathbb{E}\big[\mu_0(X)S_0(X)\big]+\mathbb{E}\Big[\mu_0(X)\underbrace{\mathbb{E}\big[S_0(A|X)|X\big]}_{=0}\Big]+\sum_{k=1}^{K}\mathbb{E}\Big[\mu_0(X)\underbrace{\mathbb{E}\big[S_0(M_k|X,A,\overline{M}_{k-1})|X,A,\overline{M}_{k-1}\big]}_{=0}\Big]\\
&\quad+\mathbb{E}\Big[\mu_0(X)\underbrace{\mathbb{E}\big[S_0(Y|X,A,\overline{M}_K)|X,A,\overline{M}_K\big]}_{=0}\Big]\\
&=\int\mu_0(x)S_0(x)dP_0(x)\\
&=\iiint yS_0(x)dP_0(y|x,a_{K+1},\overline{m}_K)\big[\prod_{k=1}^{K}dP_0(m_k|x,a_k,\overline{m}_{k-1})\big]dP_0(x)\\
&=\phi_0.
\end{aligned}
$$

Second, for $k\in[K]$,

$$
\begin{aligned}
&\mathbb{E}[\varphi_k(O)S_0(O)]\\
&=\mathbb{E}\big[\varphi_k(O)\big(S_0(X)+S_0(A|X)+\sum_{j=1}^{K}S_0(M_j|X,A,\overline{M}_{j-1})+S_0(Y|X,A,\overline{M}_K)\big)\big]\\
&=\mathbb{E}\Big[\mathbb{E}\big[\varphi_k(O)\big(S_0(X)+S_0(A|X)+\sum_{j=1}^{k-1}S_0(M_j|X,A,\overline{M}_{j-1})\big)|X,A,\overline{M}_{k-1}\big]\Big]+\mathbb{E}\big[\varphi_k(O)S_0(M_k|X,A,\overline{M}_{k-1})\big)\big]\\
&\quad+\sum_{j=k+1}^{K}\mathbb{E}\Big[\varphi_k(O)\underbrace{\mathbb{E}\big[S_0(M_j|X,A,\overline{M}_{j-1})|X,A,\overline{M}_{j-1}\big]}_{=0}\Big]+\mathbb{E}\Big[\varphi_k(O)\underbrace{\mathbb{E}\big[S_0(Y|X,A,\overline{M}_K)|X,A,\overline{M}_K\big]}_{=0}\Big]\\
&=\mathbb{E}\Big[\big(S_0(X)+S_0(A|X)+\sum_{j=1}^{k-1}S_0(M_j|X,A,\overline{M}_{j-1})\big)\underbrace{\mathbb{E}\big[\varphi_k(O)|X,A,\overline{M}_{k-1}\big]}_{=0}\Big]+\mathbb{E}\big[\varphi_k(O)S_0(M_k|X,A,\overline{M}_{k-1})\big)\big]\\
&=\mathbb{E}\big[\varphi_k(O)S_0(M_k|X,A,\overline{M}_{k-1})\big)\big]\\
&=\mathbb{E}\Big[\mathbb{E}\big[\frac{\mathbb{I}(A=a_k)}{p(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_k,\overline{M}_{j-1})}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big)S_0(M_k|X,A,\overline{M}_{k-1})\big)\big|X,A,\overline{M}_{k-1}\big]\Big]\\
&=\mathbb{E}\big[\frac{\mathbb{I}(A=a_k)}{p(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_k,\overline{M}_{j-1})}\Big)\mu_k(X,\overline{M}_k)S_0(M_k|X,A,\overline{M}_{k-1})\big)\big]\\
&=\mathbb{E}_X\mathbb{E}\big[\Big(\prod_{j=1}^{k-1}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_k,\overline{M}_{j-1})}\Big)\mu_k(X,\overline{M}_k)S_0(M_k|X,A,\overline{M}_{k-1})\big)|X,A=a_k\big]\\
&=\int_x S_0(m_k|x,a,\overline{m}_{k-1})\Big(\int_y\int_{\overline{m}_K}ydP_0(y|x,a_{K+1},\overline{m}_K)\prod_{j=k+1}^{K}dP_0(m_j|x,a_j,\overline{m}_{j-1})\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\cdot p_0(m_k|x, a_k, \overline{m}_{k-1})\Big(\prod_{j=1}^{k-1} \frac{p(m_j|x, a_j, \overline{m}_{j-1})}{p(m_j|x, a_k, \overline{m}_{j-1})}\Big)\Big(\prod_{j=1}^{k-1} p(m_j|x, a_k, \overline{m}_{j-1})\Big) dP_0(x) \\
=&\iiint y S_0(m_k|x, a, \overline{m}_{k-1}) dP_0(y|x, a_{K+1}, \overline{m}_K)\big(\prod_{j=1}^{K} dP_0(m_j|x, a_j, \overline{m}_{j-1})\big) dP_0(x) \\
=&\phi_k,
\end{aligned}
$$

where the fourth equality is due to the fact that

$$
\begin{aligned}
&\mathbb{E}\big[\varphi_k(O)|X, A, \overline{M}_{k-1}\big] \\
=&\mathbb{E}\big[\frac{\mathbb{I}(A = a_k)}{p(a_k|X)}\Big(\prod_{j=1}^{k-1} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_k, \overline{M}_{j-1})}\Big)\big(\mu_k(X, \overline{M}_k) - \mu_{k-1}(X, \overline{M}_{k-1})\big)|X, A, \overline{M}_{k-1}\big] \\
=&\mathbb{E}\big[\Big(\prod_{j=1}^{k-1} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_k, \overline{M}_{j-1})}\Big)\big(\mu_k(X, \overline{M}_k) - \mu_{k-1}(X, \overline{M}_{k-1})\big)|X, A = a_k, \overline{M}_{k-1}\big] \\
=&\Big(\prod_{j=1}^{k-1} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_k, \overline{M}_{j-1})}\Big) \underbrace{\mathbb{E}\big[\mu_k(X, \overline{M}_k) - \mu_{k-1}(X, \overline{M}_{k-1})|X, A = a_k, \overline{M}_{k-1}\big]}_{=0} \\
=&0.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\mathbb{E}[\varphi_{K+1}(O)S_0(O)] \\
=&\mathbb{E}\big[\varphi_{K+1}(O)\big(S_0(X) + S_0(A|X) + \sum_{j=1}^{K} S_0(M_j|X, A, \overline{M}_{j-1}) + S_0(Y|X, A, \overline{M}_K)\big)\big] \\
=&\mathbb{E}\Big[\mathbb{E}\big[\varphi_{K+1}(O)\big(S_0(X) + S_0(A|X) + \sum_{j=1}^{K} S_0(M_j|X, A, \overline{M}_{j-1})\big)|X, A, \overline{M}_K\big]\Big] + \mathbb{E}\big[\varphi_{K+1}(O)S_0(Y|X, A, \overline{M}_K)\big)\big] \\
=&\mathbb{E}\Big[\big(S_0(X) + S_0(A|X) + \sum_{j=1}^{K} S_0(M_j|X, A, \overline{M}_{j-1})\big) \underbrace{\mathbb{E}\big[\varphi_{K+1}(O)|X, A, \overline{M}_K\big]}_{=0}\Big] + \mathbb{E}\big[\varphi_{K+1}(O)S_0(Y|X, A, \overline{M}_K)\big)\big] \\
=&\mathbb{E}\big[\varphi_{K+1}(O)S_0(Y|X, A, \overline{M}_K)\big)\big] \\
=&\mathbb{E}\Big[\mathbb{E}\big[\frac{\mathbb{I}(A = a_{K+1})}{p(a_{K+1}|X)}\Big(\prod_{j=1}^{K} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_{K+1}, \overline{M}_{j-1})}\Big)\big(Y - \mu_K(X, \overline{M}_K)\big)S_0(Y|X, A, \overline{M}_K)\big)\big|X, A, \overline{M}_K\big]\Big] \\
=&\mathbb{E}\Big[\mathbb{E}\big[\frac{\mathbb{I}(A = a_{K+1})}{p(a_{K+1}|X)}\Big(\prod_{j=1}^{K} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_{K+1}, \overline{M}_{j-1})}\Big) Y S_0(Y|X, A, \overline{M}_K)\big)\big|X, A, \overline{M}_K\big]\Big] \\
=&\mathbb{E}\Big[\mathbb{E}\big[\frac{\mathbb{I}(A = a_{K+1})}{p(a_{K+1}|X)}\Big(\prod_{j=1}^{K} \frac{p(M_j|X, a_j, \overline{M}_{j-1})}{p(M_j|X, a_{K+1}, \overline{M}_{j-1})}\Big) Y S_0(Y|X, A, \overline{M}_K)\big)\big|X, A\big]\Big]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}_X\Big[\mathbb{E}\big[\Big(\prod_{j=1}^{K}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)YS_0(Y|X,A,\overline{M}_K)\big)\big|X,A=a_{K+1}\big]\Big]\\
=&\iiint yS_0(y|x,a,\overline{m}_K)dP_0(y|x,a_{K+1},\overline{m}_K)\Big(\prod_{j=1}^{K}\frac{p(m_j|x,a_j,\overline{m}_{j-1})}{p(m_j|x,a_{K+1},\overline{m}_{j-1})}\Big)\big(\prod_{j=1}^{K}dP_0(m_j|x,a_{K+1},\overline{m}_{j-1})\big)dP_0(x)\\
=&\iiint yS_0(y|x,a,\overline{m}_K)dP_0(y|x,a_{K+1},\overline{m}_K)\big(\prod_{j=1}^{K}dP_0(m_j|x,a_j,\overline{m}_{j-1})\big)dP_0(x)\\
=&\phi_{K+1},
\end{aligned}
$$

where the third equality is due to the fact that

$$
\begin{aligned}
&\mathbb{E}\big[\varphi_{K+1}(O)|X,A,\overline{M}_K\big]\\
=&\mathbb{E}\big[\frac{\mathbb{I}(A=a_{K+1})}{p(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big)|X,A,\overline{M}_K\big]\\
=&\frac{p(a_{K+1}|X,\overline{M}_K)}{p(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{p(M_j|X,a_j,\overline{M}_{j-1})}{p(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)|X,A=a_{K+1},\overline{M}_K\big]}_{=0}\\
=&0.
\end{aligned}
$$

Since $\phi_k=\mathbb{E}[\varphi_k(O)S_0(O)]$ for all $k\in\{0,\dots K+1\}$, we have

$$
\frac{\partial\psi_{\overline{a}}(t)}{\partial t}\Big|_{t=0}=\sum_{k=0}^{K+1}\phi_k=\mathbb{E}\big[\big(\sum_{k=0}^{K+1}\varphi_k(O)\big)S_0(O)\big]=\mathbb{E}\big[\varphi_{\overline{a}}(O)S_0(O)\big].
$$

D Proof of Theorems 3 and 4

D.1 Parametric Estimation of Nuisance Parameters

In this subsection, we prove the multiple robustness of $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ and $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ for the case where parametric models are used to estimate the corresponding nuisance functions. The local efficiency of these estimators is implied by our proof in section D.2, which considers the case where data-adaptive methods and cross-fitting are used to estimate the nuisance functions.

Let us start with $\hat{\psi}_{\overline{a}}^{\text{eif}_1}=\mathbb{P}_n[m_1(O;\hat{\eta}_1)]$, where $m_1(O;\hat{\eta}_1)$ denotes the quantity inside $\mathbb{P}_n[\cdot]$ in equation (12), and $\hat{\eta}_1=(\hat{\pi}_0,\hat{f}_1,\dots\hat{f}_K,\hat{\mu}_K)$. In the meantime, let $\eta_1=(\pi_0,f_1,\dots f_K,\mu_K)$ denote the truth and $\eta_1^*=(\pi_0^*,f_1^*,\dots f_K^*,\mu_K^*)$ the probability limit of $\hat{\eta}_1$. A first-order Taylor expansion of $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ yields

$$
\hat{\psi}_{\overline{a}}^{\text{eif}_1}=\mathbb{P}_n\big[m_1(O;\eta_1^*)\big]+o_p(1).
$$

Hence it suffices to show $\mathbb{E}[m_1(O;\eta_1^*)]=\psi_{\overline{a}}$ whenever all but one elements in η_1^* equal the truth. Consistency follows from the law of large numbers. By treating $\hat{\psi}_{\overline{a}}^{\text{eif}_1}=\mathbb{P}_n[m_1(O;\hat{\eta}_1)]$ as a twostage M-estimator, asymptotic normality follows from standard regularity conditions for estimating equations (e.g., Newey and McFadden 1994, p. 2148).

First, if $\eta_1^* = (\pi_0^*, f_1, \dots f_K, \mu_K)$, the MLE of μ_k $(0 \leq k \leq K-1)$ will also be consistent. Thus,

$$
\begin{aligned}
&\mathbb{E}[m_1(O;\eta_1^*)] \\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0^*(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big) \\
&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0^*(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big) \\
&+\mu_0(X)\Big] \\
=&\mathbb{E}\Big[\frac{\pi_0(a_{K+1}|X,\overline{M}_K)}{\pi_0^*(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)\big|X,A=a_{K+1},\overline{M}_K\big]}_{=0} \\
&+\sum_{k=1}^{K}\frac{\pi_0(a_k|X,\overline{M}_{k-1})}{\pi_0^*(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}\Big)\underbrace{\mathbb{E}\big[\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big|X,A=a_k,\overline{M}_{k-1}\big]}_{=0} \\
&+\mu_0(X)\Big] \\
=&\mathbb{E}[\mu_0(X)] \\
=&\psi_{\overline{a}}.
\end{aligned}
$$

Second, if $\eta_1^* = (\pi_0, f_1, \dots f_{k'-1}, f_{k'}^*, f_{k'+1}, \dots f_K, \mu_K)$, the MLE of μ_k for any $k \geq k'$ will also be consistent. Thus,

$$
\begin{aligned}
&\mathbb{E}[m_1(O;\eta_1^*)] \\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j^*(M_j|X,a_j,\overline{M}_{j-1})}{f_j^*(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big) \\
&+\sum_{k=k'+1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{f_j^*(M_j|X,a_j,\overline{M}_{j-1})}{f_j^*(M_j|X,a_k,\overline{M}_{j-1})}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big) \\
&+\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_{k'}|X)}\Big(\prod_{j=1}^{k'-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)\big(\mu_{k'}(X,\overline{M}_{k'})-\mu_{k'-1}^*(X,\overline{M}_{k'-1})\big) \\
&+\sum_{k=1}^{k'-1}\frac{\mathbb{I}(A=a_k)}{\pi_0(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}\Big)\big(\mu_k^*(X,\overline{M}_k)-\mu_{k-1}^*(X,\overline{M}_{k-1})\big) \\
&+\mu_0^*(X)\Big] \\
=&\mathbb{E}\Big[\frac{\pi_0(a_{K+1}|X,\overline{M}_K)}{\pi_0(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j^*(M_j|X,a_j,\overline{M}_{j-1})}{f_j^*(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)\big|X,A=a_{K+1},\overline{M}_K\big]}_{=0}
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{k=k'+1}^{K} \frac{\pi_0(a_k|X,\overline{M}_{k-1})}{\pi_0(a_k|X)}\Big(\prod_{j=1}^{k-1}\frac{f_j^*(M_j|X,a_j,\overline{M}_{j-1})}{f_j^*(M_j|X,a_k,\overline{M}_{j-1})}\Big)\underbrace{\mathbb{E}\big[\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big|X,A=a_k,\overline{M}_{k-1}\big]}_{=0}\\
&+ \frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_{k'}|X)}\Big(\prod_{j=1}^{k'-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)\mu_{k'}(X,\overline{M}_{k'})\\
&+ \sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\mathbb{E}\Big[\Big(\frac{\mathbb{I}(A=a_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}-\frac{\mathbb{I}(A=a_{k+1})}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k+1},\overline{M}_{j-1})}\Big)|X,\overline{M}_k\Big]\\
&+ \mu_0^*(X)\underbrace{\mathbb{E}\big[1-\frac{\mathbb{I}(A=a_1)}{\pi_0(a_1|X)}|X\big]}_{=0}\Big]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_{k'}|X)}\Big(\prod_{j=1}^{k'-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)\mu_{k'}(X,\overline{M}_{k'})\Big]\\
&+\mathbb{E}\Big[\sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\Big(\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}-\frac{\pi_k(a_{k+1}|X,\overline{M}_k)}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k+1},\overline{M}_{j-1})}\Big)\Big]\\
=&\underbrace{\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_{k'}|X)}\Big(\prod_{j=1}^{k'-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)\mu_{k'}(X,\overline{M}_{k'})\Big]}_{=\psi_{\overline{a}}}\\
&+\mathbb{E}\Big[\sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\underbrace{\Big(\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}-\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)}_{=0}\Big]\\
=&\psi_{\overline{a}},
\end{aligned}
$$

where the penultimate equality is due to the fact that

$$
\begin{aligned}
&\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}\\
=&\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\Big(\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_k|X,\overline{M}_j)}\cdot\frac{\pi_{j-1}(a_k|X,\overline{M}_{j-1})}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)\\
=&\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\Big(\frac{\pi_{j-1}(a_k|X,\overline{M}_{j-1})}{\pi_j(a_k|X,\overline{M}_j)}\Big)\prod_{j=1}^{k-1}\Big(\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)\\
=&\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_{k-1}(a_k|X,\overline{M}_{k-1})}\prod_{j=1}^{k-1}\Big(\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)\\
=&\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}
\end{aligned}
$$

and that

$$
\begin{aligned}
&\frac{\pi_k(a_{k+1}|X,\overline{M}_k)}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k+1},\overline{M}_{j-1})}\\
=&\frac{\pi_k(a_{k+1}|X,\overline{M}_k)}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\Big(\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{k+1}|X,\overline{M}_j)}\cdot\frac{\pi_{j-1}(a_{k+1}|X,\overline{M}_{j-1})}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)\\
=&\frac{\pi_k(a_{k+1}|X,\overline{M}_k)}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\Big(\frac{\pi_{j-1}(a_{k+1}|X,\overline{M}_{j-1})}{\pi_j(a_{k+1}|X,\overline{M}_j)}\Big)\prod_{j=1}^{k}\Big(\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)\\
=&\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}.
\end{aligned}
$$

Finally, if $\eta_1^*=(\pi_0,f_1,\dots f_K,\mu_K^*)$, we have

$$
\begin{aligned}
&\mathbb{E}[m_1(O;\eta_1^*)]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{K+1},\overline{M}_{j-1})}\Big)\big(Y-\mu_K^*(X,\overline{M}_K)\big)\\
&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0(a_k|X)}\Big(\prod_{j=1}^{k}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}\Big)\big(\mu_k^*(X,\overline{M}_k)-\mu_{k-1}^*(X,\overline{M}_{k-1})\big)+\mu_0^*(X)\Big]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)Y\\
&+\sum_{k=1}^{K}\mu_k^*(X,\overline{M}_k)\underbrace{\mathbb{E}\Big[\Big(\frac{\mathbb{I}(A=a_k)}{\pi_0(a_k|X)}\prod_{j=1}^{k-1}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_k,\overline{M}_{j-1})}-\frac{\mathbb{I}(A=a_{k+1})}{\pi_0(a_{k+1}|X)}\prod_{j=1}^{k}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k+1},\overline{M}_{j-1})}\Big)|X,\overline{M}_k\Big]}_{=0\quad\text{(same as the previous case)}}\\
&+\mu_0^*(X)\underbrace{\mathbb{E}\Big[1-\frac{\mathbb{I}(A=a_1)}{\pi_0(a_1|X)}||X\Big]}_{=0}\Big]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_{K+1}|X)}\Big(\prod_{j=1}^{K}\frac{f_j(M_j|X,a_j,\overline{M}_{j-1})}{f_j(M_j|X,a_{k'},\overline{M}_{j-1})}\Big)Y\Big]\\
=&\psi_{\overline{a}}.
\end{aligned}
$$

Now consider $\hat{\psi}_{\overline{a}}^{\text{eif}_2}=\mathbb{P}_n[m_2(O;\hat{\eta}_2)]$, where $m_2(O;\hat{\eta}_2)$ denotes the quantity inside $\mathbb{P}_n[\cdot]$ in equation (14), and $\hat{\eta}_2=(\hat{\pi}_0,\dots\hat{\pi}_K,\hat{\mu}_0,\dots\hat{\mu}_K)$. In the meantime, let $\eta_2=(\pi_0,\dots\pi_K,\mu_0,\dots\mu_K)$ denote the truth and $\eta_2^*=(\pi_0^*,\dots\pi_K^*,\mu_0^*,\dots\mu_K^*)$ denote the probability limit of $\hat{\eta}_2$. A first-order Taylor expansion of $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ yields

$$\hat{\psi}_{\overline{a}}^{\text{eif}_2}=\mathbb{P}_n\big[m_2(O;\eta_2^*)\big]+o_p(1).$$

Hence it suffices to show $\mathbb{E}[m_2(O;\eta_2^*)] = \psi_{\overline{a}}$ if

$$\eta_2^* = (\pi_0, \dots \pi_{k'-1}, \pi_{k'}^*, \dots \pi_K^*, \mu_0^*, \dots \mu_{k'-1}^*, \mu_{k'}, \dots \mu_K)$$

for any $k' \in \{0, \dots K+1\}$.

First, it $k' = 0$, then all the outcome models are correctly specified, which implies

$$\begin{aligned}
&\mathbb{E}[m_2(O;\eta_2^*)] \\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big) \\
&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big) \\
&+\mu_0(X)\Big] \\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)\big|X,A=a_{K+1},\overline{M}_K\big]}_{=0} \\
&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big|X,A=a_k,\overline{M}_{k-1}\big]}_{=0} \\
&+\mu_0(X)\Big] \\
=&\mathbb{E}[\mu_0(X)] \\
=&\psi_{\overline{a}}.
\end{aligned}$$

Second, if $k' \in \{1, \dots K-1\}$, we have

$$\begin{aligned}
&\mathbb{E}[m_2(O;\eta_2^*)] \\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big) \\
&+\sum_{k=k'+1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big) \\
&+\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k'-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_{k'}(X,\overline{M}_{k'})-\mu_{k'-1}^*(X,\overline{M}_{k'-1})\big) \\
&+\sum_{k=1}^{k'-1}\frac{\mathbb{I}(A=a_k)}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_k^*(X,\overline{M}_k)-\mu_{k-1}^*(X,\overline{M}_{k-1})\big) \\
&+\mu_0^*(X)\Big]
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)\big|X,A=a_{K+1},\overline{M}_K\big]}_{=0}\\
&+\sum_{k=k'+1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0^*(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j^*(a_j|X,\overline{M}_j)}{\pi_j^*(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big|X,A=a_k,\overline{M}_{k-1}\big]}_{=0}\\
&+\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k'-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\mu_{k'}(X,\overline{M}_{k'})\\
&+\sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\mathbb{E}\big[\frac{\mathbb{I}(A=a_k)}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)-\frac{\mathbb{I}(A=a_{k+1})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)|X,\overline{M}_k\big]\\
&+\mu_0^*(X)\underbrace{\mathbb{E}\big[1-\frac{\mathbb{I}(A=a_1)}{\pi_0(a_1|X)}|X\big]}_{=0}\Big]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k'-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\mu_{k'}(X,\overline{M}_{k'})\Big]\\
&+\mathbb{E}\Big[\sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\Big(\frac{\pi_k(a_k|X,\overline{M}_k)}{\pi_0(a_1|X)}\prod_{j=1}^{k-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}-\frac{\pi_k(a_{k+1}|X,\overline{M}_k)}{\pi_0(a_1|X)}\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\Big]\\
=&\underbrace{\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{k'})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k'-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\mu_{k'}(X,\overline{M}_{k'})\Big]}_{=\psi_{\overline{a}}}\\
&+\mathbb{E}\Big[\sum_{k=1}^{k'-1}\mu_k^*(X,\overline{M}_k)\underbrace{\Big(\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}-\prod_{j=1}^{k}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_{j-1}(a_j|X,\overline{M}_{j-1})}\Big)}_{=0}\Big]\\
=&\psi_{\overline{a}}.
\end{aligned}
$$

Finally, if $k'=K$, we have

$$
\begin{aligned}
&\mathbb{E}[m_2(O;\eta_2^*)]\\
=&\mathbb{E}\big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(Y-\mu_K^*(X,\overline{M}_K)\big)\\
&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_k^*(X,\overline{M}_k)-\mu_{k-1}^*(X,\overline{M}_{k-1})\big)+\mu_0^*(X)\big]\\
=&\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)Y
\end{aligned}
$$

$$+\sum_{k=1}^{K}\mu_k^*(X,\overline{M}_k)\underbrace{\mathbb{E}\Big[\Big(\frac{\mathbb{I}(A=a_k)}{\pi_0(a_1|X)}\prod_{j=1}^{k-1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}-\frac{\mathbb{I}(A=a_{k+1})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{k+1}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)|X,\overline{M}_k\Big]}_{=0 \quad \text{(same as the previous case)}}$$

$$+\mu_0^*(X)\underbrace{\mathbb{E}\Big[1-\frac{\mathbb{I}(A=a_1)}{\pi_0(a_1|X)}||X\Big]}_{=0}\Big]$$

$$=\mathbb{E}\Big[\frac{\mathbb{I}(A=a_{K+1})}{\pi_0(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\pi_j(a_j|X,\overline{M}_j)}{\pi_j(a_{j+1}|X,\overline{M}_j)}\Big)Y\Big]$$

$$=\psi_{\overline{a}}.$$

D.2 Data-Adaptive Estimation of Nuisance Parameters

Let us start with $\hat{\psi}_{\overline{a}}^{\text{eif}_2}=\mathbb{P}_n[m_2(O;\hat{\eta}_2)]$. Let $\tilde{\eta}_2=(\hat{\pi}_0,\dots\hat{\pi}_K,\mu_0,\dots\mu_K)$ denote a combination of estimated treatment models $\hat{\pi}_j$ and true outcome models μ_j $(0\le j\le K+1)$, and let $Pg=\int gdP$ denote the expectation of a function g of observed data O at the true model P. $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ can now be written as

$$\begin{aligned}&\hat{\psi}_{\overline{a}}^{\text{eif}_2}-\psi_{\overline{a}}\\&=\mathbb{P}_n[m_2(O;\hat{\eta}_2)]-P[m_2(O;\eta_2)]\\&=\mathbb{P}_n[\underbrace{m_2(O;\eta_2)-\psi_{\overline{a}}}_{=\varphi_{\overline{a}}(O)}]+\underbrace{P[m_2(O;\hat{\eta}_2)-m_2(O;\eta_2)]}_{\triangleq R_2}+(\mathbb{P}_n-P)[m_2(O;\hat{\eta}_2)-m_2(O;\eta_2)],\end{aligned}\tag{25}$$

The first term in equation (25) is a sample average of the efficient influence function and has an asymptotic variance of $\mathbb{E}[\big(\varphi_{\overline{a}}(O)\big)^2]$. The last term is an empirical process term that will be $o_p(n^{-1/2})$ either when parametric models are used to estimate the nuisance functions or when cross-fitting is used to induce independence between $\hat{\eta}_2$ and O (Chernozhukov *et al.* 2018). Thus it remains to analyze the second term R_2. We first observe that

$$\begin{aligned}P[m_2(O;\tilde{\eta}_2)]=&P\Big[\frac{\mathbb{I}(A=a_{K+1})}{\hat{\pi}_0(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(Y-\mu_K(X,\overline{M}_K)\big)\\&+\sum_{k=1}^{K}\frac{\mathbb{I}(A=a_k)}{\hat{\pi}_0(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\big(\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big)+\mu_0(X)\Big]\\&=P\Big[\frac{\pi_0(a_{K+1}|X,\overline{M}_K)}{\hat{\pi}_0(a_1|X)}\Big(\prod_{j=1}^{K}\frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[Y-\mu_K(X,\overline{M}_K)\big|X,A=a_{K+1},\overline{M}_K\big]}_{=0}\\&+\sum_{k=1}^{K}\frac{\pi_0(a_k|X,\overline{M}_{k-1})}{\hat{\pi}_0(a_1|X)}\Big(\prod_{j=1}^{k-1}\frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\underbrace{\mathbb{E}\big[\mu_k(X,\overline{M}_k)-\mu_{k-1}(X,\overline{M}_{k-1})\big|X,A=a_k,\overline{M}_{k-1}\big]}_{=0}\\&+\mu_0(X)\Big]\end{aligned}$$

$$= P[\mu_0(X)]$$
$$= P[m_2(O;\eta_2)].$$

Then, by substituting $m_2(O;\tilde{\eta}_2)$ for $m_2(O;\eta_2)$ in R_2, rearranging terms, and applying the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
R_2 &= P[m_2(O;\hat{\eta}_2) - m_2(O;\tilde{\eta}_2)] \\
&= P\Big[\frac{\big(\hat{\pi}_0(a_1|X) - \pi_0(a_1|X)\big)\big(\hat{\mu}_0(X) - \mu_0(X)\big)}{\hat{\pi}_0(a_1|X)}\Big] \\
&+ \sum_{k=1}^{K} P\Big[\Big(\prod_{j=1}^{k} \frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\frac{\big(\hat{\pi}_k(a_{k+1}|X,\overline{M}_k) - \pi_k(a_{k+1}|X,\overline{M}_k)\big)\big(\hat{\mu}_k(X,\overline{M}_k) - \mu_k(X,\overline{M}_k)\big)}{\hat{\pi}_0(a_1|X)}\Big] \\
&- \sum_{k=1}^{K} P\Big[\Big(\prod_{j=1}^{k-1} \frac{\hat{\pi}_j(a_j|X,\overline{M}_j)}{\hat{\pi}_j(a_{j+1}|X,\overline{M}_j)}\Big)\frac{\big(\hat{\pi}_k(a_{k}|X,\overline{M}_k) - \pi_k(a_{k}|X,\overline{M}_k)\big)\big(\hat{\mu}_k(X,\overline{M}_k) - \mu_k(X,\overline{M}_k)\big)}{\hat{\pi}_0(a_1|X)}\Big] \\
&= \sum_{k=0}^{K} O_p\big(\|\hat{\pi}_k(a_{k+1}|X,\overline{M}_k) - \pi_k(a_{k+1}|X,\overline{M}_k)\| \cdot \|\hat{\mu}_k(X,\overline{M}_k) - \mu_k(X,\overline{M}_k)\|\big) \\
&+ \sum_{k=1}^{K} O_p\big(\|\hat{\pi}_k(a_{k}|X,\overline{M}_k) - \pi_k(a_{k}|X,\overline{M}_k)\| \cdot \|\hat{\mu}_k(X,\overline{M}_k) - \mu_k(X,\overline{M}_k)\|\big) \qquad (26)
\end{aligned}$$

where $\|g\| = (\int g^T g dP)^{1/2}$. The last equality uses the positivity assumption that $\pi_k(a|X,\overline{M}_k)$ is bounded away from zero for any k and any a. Thus, assuming that the empirical process term is on the order of $o_p(n^{-1/2})$ (e.g., via cross-fitting), we can write equation (25) as

$$\hat{\psi}_{\overline{a}}^{\text{eif}_2} - \psi_{\overline{a}} = \mathbb{P}_n[m_2(O;\eta_2) - \psi_{\overline{a}}] + \sum_{k=0}^{K} O_p(\|\hat{\pi}_k - \pi_k\|) \cdot O_p(\|\hat{\mu}_k - \mu_k\|) + o_p(n^{-1/2}),$$

where $\pi_k = (\pi_k(0|X,\overline{M}_k),\, \pi_k(1|X,\overline{M}_k))^T$. Therefore, $\hat{\psi}_{\overline{a}}^{\text{eif}_2}$ is consistent if $\sum_{k=0}^{K} O_p(\|\hat{\pi}_k - \pi_k\|) \cdot O_p(\|\hat{\mu}_k - \mu_k\|) = o_p(1)$, and it is semiparametric efficient if $\sum_{k=0}^{K} O_p(\|\hat{\pi}_k - \pi_k\|) \cdot O_p(\|\hat{\mu}_k - \mu_k\|) = o_p(n^{-1/2})$.

Now let us consider $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$. In a similar vein, we can write $\hat{\psi}_{\overline{a}}^{\text{eif}_1} - \psi_{\overline{a}}$ as

$$\hat{\psi}_{\overline{a}}^{\text{eif}_1} - \psi_{\overline{a}} = \mathbb{P}_n[m_1(O;\eta_1) - \psi_{\overline{a}}] + \sum_{k=0}^{K} O_p(\|\check{\pi}_k - \pi_k\|) \cdot O_p(\|\check{\mu}_k - \mu_k\|) + o_p(n^{-1/2}),$$

where $\check{\pi}_k$ and $\check{\mu}_k$ are estimates of π_k and μ_k constructed from $\hat{\eta}_1 = \{\hat{\pi}_0, \hat{f}_1, \dots \hat{f}_K, \hat{\mu}_K\}$. We first note that for any a, $\check{\pi}_k(a|X,\overline{M}_k) - \pi_k(a|X,\overline{M}_k)$ can be decomposed as

$$\begin{aligned}
&\check{\pi}_k(a|X,\overline{M}_k) - \pi_k(a|X,\overline{M}_k) \\
=&\frac{\hat{p}(\overline{M}_k|X,a)\hat{\pi}_0(a|X)}{\sum_{a'}\hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)} - \frac{p(\overline{M}_k|X,a)\pi_0(a|X)}{\sum_{a'}p(\overline{M}_k|X,a')\pi_0(a'|X)}
\end{aligned}$$

$$= \underbrace{\frac{\hat{p}(\overline{M}_k|X,a)\big(\hat{\pi}_0(a|X) - \pi_0(a|X)\big)}{\sum_{a'} \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)}}_{\triangleq \Delta_\pi^1} + \underbrace{\frac{\big(\hat{p}(\overline{M}_k|X,a) - p(\overline{M}_k|X,a)\big)\pi_0(a|X)}{\sum_{a'} \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)}}_{\triangleq \Delta_\pi^2} +$$
$$\underbrace{\frac{p(\overline{M}_k|X,a)\pi_0(a|X)\sum_{a'}\big(p(\overline{M}_k|X,a')\pi_0(a'|X) - \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)\big)}{\sum_{a'} \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)\sum_{a'} p(\overline{M}_k|X,a')\pi_0(a'|X)}}_{\triangleq \Delta_\pi^3}.$$

By the positivity assumption, we know that $\|\Delta_\pi^1\| = O_p(\|\hat{\pi}_0 - \pi_0\|)$. Using the factorization $p(\overline{M}_k|X,a) = \prod_{j=1}^k p(M_j|X,a,\overline{M}_{j-1})$, $\|\Delta_\pi^2\|$ can be expressed as

$$\begin{aligned}
\|\Delta_\pi^2\| &= \Big\| \frac{\pi_0(a|X)\big(\prod_{j=1}^k \hat{f}_j(M_j|X,a,\overline{M}_{j-1}) - \prod_{j=1}^k f_j(M_j|X,a,\overline{M}_{j-1})\big)}{\sum_{a'} \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)} \Big\| \\
&= \Big\| \frac{\pi_0(a|X)}{\sum_{a'} \hat{p}(\overline{M}_k|X,a')\hat{\pi}_0(a'|X)} \cdot \\
&\sum_{l=1}^k \big(\prod_{j=1}^{l-1} \hat{f}_j(M_j|X,a,\overline{M}_{j-1}) \prod_{j=l+1}^k f_j(M_j|X,a,\overline{M}_{j-1})\big)\big(\hat{f}_l(M_l|X,a,\overline{M}_{l-1}) - f_l(M_l|X,a,\overline{M}_{l-1})\big) \Big\| \\
&= \sum_{l=1}^k O_p(\|\hat{f}_l - f_l\|),
\end{aligned}$$

where $f_l = (f_l(M_l|X,0,\overline{M}_{l-1}),\ f_l(M_l|X,1,\overline{M}_{l-1}))^T$. By a similar logic, $\|\Delta_\pi^3\|$ can be written as

$$\|\Delta_\pi^3\| = O_p(\|\hat{\pi}_0 - \pi_0\|) + \sum_{l=1}^k O_p(\|\hat{f}_l - f_l\|).$$

In sum, we have

$$\|\breve{\pi}_k - \pi_k\| = O_p(\|\hat{\pi}_0 - \pi_0\|) + \sum_{l=1}^k O_p(\|\hat{f}_l - f_l\|). \tag{27}$$

Now consider $\|\breve{\mu}_k - \mu_k\|$. Using the fact that

$$\mu_k(x,\overline{m}_k) = \int \mu_K(x,\overline{m}_k)\big(\prod_{j=k+1}^K p(m_j|x,a_j,\overline{m}_{j-1})dm_j\big),$$

we can decompose $\breve{\mu}_k(x,\overline{m}_k) - \mu_k(x,\overline{m}_k)$ into

$$\begin{aligned}
&\breve{\mu}_k(x,\overline{m}_k) - \mu_k(x,\overline{m}_k) \\
&= \int \big(\hat{\mu}_K(x,\overline{m}_K) - \mu_K(x,\overline{m}_K)\big)\big(\prod_{j=k+1}^K \hat{f}_j(m_j|x,a_j,\overline{m}_{j-1})dm_j\big)
\end{aligned}$$

$$+\sum_{l=k+1}^{K}\int \mu_K(x,\overline{m}_K)\big(\big(\hat{f}_l(m_l|x,a_l,\overline{m}_{l-1})-f_l(m_l|x,a_l,\overline{m}_{l-1})\big)dm_l\big)\cdot$$

$$\big(\prod_{j=k+1}^{l-1}\hat{f}_j(m_j|x,a_j,\overline{m}_{j-1})dm_j\big)\big(\prod_{j=l+1}^{K}f_j(m_j|x,a_j,\overline{m}_{j-1})dm_j\big)$$

$$=\int\big(\hat{\mu}_K(x,\overline{m}_K)-\mu_K(x,\overline{m}_K)\big)\underbrace{\Big(\frac{\prod_{j=k+1}^{K}\hat{f}_j(m_j|x,a_j,\overline{m}_{j-1})}{\prod_{j=k+1}^{K}f_j(m_j|x,\overline{m}_{j-1})}\Big)}_{\triangleq g(x,\overline{m}_K)}\big(\prod_{j=k+1}^{K}f_j(m_j|x,\overline{m}_{j-1})dm_j\big)$$

$$+\sum_{l=k+1}^{K}\int\big(\hat{f}_l(m_l|x,a_l,\overline{m}_{l-1})-f_l(m_l|x,a_l,\overline{m}_{l-1})\big)$$

$$\underbrace{\frac{\mu_K(x,\overline{m}_K)\big(\prod_{j=k+1}^{l-1}\hat{f}_j(m_j|x,a_j,\overline{m}_{j-1})\big)}{\prod_{j=k+1}^{l}f_j(m_j|x,\overline{m}_{j-1})}}_{\triangleq h_l(x,\overline{m}_K)}\big(\prod_{j=k+1}^{K}f_j(m_j|x,\overline{m}_{j-1})dm_j\big)$$

Using the notation $dP_2=\prod_{j=k+1}^{K}f_j(m_j|x,\overline{m}_{j-1})dm_j$ and $dP_1=dP_X(x)\cdot\prod_{j=1}^{k}f_j(m_j|x,\overline{m}_{j-1})dm_j$, we have

$$\begin{aligned}
&\|\check{\mu}_k-\mu_k\|\\
=&\|\int(\hat{\mu}_K-\mu_K)g\,dP_2+\sum_{l=k+1}^{K}\int(\hat{f}_l-f_l)h_l\,dP_2\|_{P_1}\\
\le&\|\int(\hat{\mu}_K-\mu_K)g\,dP_2\|_{P_1}+\sum_{l=k+1}^{K}\|\int(\hat{f}_l-f_l)h_l\,dP_2\|_{P_1}\\
=&\big[\int\big(\int(\hat{\mu}_K-\mu_K)g\,dP_2\big)^2dP_1\big]^{1/2}+\sum_{l=k+1}^{K}\big[\int\big(\int(\hat{f}_l-f_l)h_l\,dP_2\big)^2dP_1\big]^{1/2}\\
\le&\big[\int\big(\int(\hat{\mu}_K-\mu_K)^2dP_2\big)\big(\int g^2dP_2\big)dP_1\big]^{1/2}+\sum_{l=k+1}^{K}\big[\int\big(\int(\hat{f}_l-f_l)^2dP_2\big)\big(\int h_l^2dP_2\big)dP_1\big]^{1/2}\ \text{(Cauchy-Schwartz)}\\
\le&\big[\int(\hat{\mu}_K-\mu_K)^2dP_2dP_1\cdot\|\int g^2dP_2\|_{P_1,\infty}\big]^{1/2}+\sum_{l=k+1}^{K}\big[\int(\hat{f}_l-f_l)^2dP_2dP_1\cdot\|\int h_l^2dP_2\|_{P_1,\infty}\big]^{1/2}
\end{aligned}$$

Under the positivity assumption and the assumption that $\mu_K(X,\overline{M}_K)$ is bounded with probability one, we have $\|\int g^2dP_2\|_{P_1,\infty}<\infty$ and $\|\int h_l^2dP_2\|_{P_1,\infty}<\infty$. Thus,

$$\|\check{\mu}_k-\mu_k\|=O_p(\|\hat{\mu}_K-\mu_K\|)+\sum_{l=k+1}^{K}O_p(\|\hat{f}_l-f_l\|)\tag{28}$$

From equations (27-28), we have

$$\sum_{k=0}^{K} O_p(\|\check{\pi}_k - \pi_k\|) \cdot O_p(\|\check{\mu}_k - \mu_k\|)$$
$$= \sum_{k=0}^{K} \big(O_p(\|\hat{\pi}_0 - \pi_0\|) + \sum_{l=1}^{k} O_p(\|\hat{f}_l - f_l\|)\big)\big(O_p(\|\hat{\mu}_K - \mu_K\|) + \sum_{l=k+1}^{K} O_p(\|\hat{f}_l - f_l\|)\big),$$

which implies that $\hat{\psi}_{\overline{a}}^{\text{eif}_1}$ is consistent if $\sum_{u,v \in \hat{\eta}; u \neq v} R_n(u) R_n(v) = o(1)$, and semiparametric efficient if $\sum_{u,v \in \hat{\eta}; u \neq v} R_n(u) R_n(v) = o(n^{-1/2})$.

E More Details of the Simulation Study

The variables X, A, M_1, M_2, Y in the simulation study are generated via the following model:

$$\begin{aligned}
(U_{XA}, U_{XM_1}, U_{XM_2}, U_{XY}) &\sim N(0, I_4) \\
X &\sim N((U_{XA}, U_{XM_1}, U_{XM_2}, U_{XY})\beta_X, 1) \\
A &\sim \text{Bernoulli}\big(\text{logit}^{-1}[(1, U_{XA}, |X|)\beta_A]\big) \\
M_1 &\sim N\big((1, U_{XM_1}, X, X^2, A)\beta_{M_1}, 1\big) \\
M_2 &\sim N\big((1, U_{XM_2}, X, X^2, A, M_1, XA, XM_1)\beta_{M_2}, 1\big) \\
Y &\sim N\big((1, U_{XY}, X, X^2, A, M_1, XM_1, M_2, AM_2)\beta_Y, 1\big).
\end{aligned}$$

The coefficients $\beta_X, \beta_A, \beta_{M_1}, \beta_{M_2}, \beta_Y$ are generated using a set of uniform distributions with certain constraints designed to create nontrivial degrees of model misspecification.

It can be shown that under the above model, the 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 $\mu_2(x, m_1, m_2)$ can be consistently estimated via the following GLMs:

$$\begin{aligned}
\pi_0(1|X) &= \text{logit}^{-1}[(1, |X|)\gamma_0] \\
\pi_1(1|X, M_1) &= \text{logit}^{-1}[(1, X, X^2, |X|, M_1)\gamma_1] \\
\pi_2(1|X, M_1, M_2) &= \text{logit}^{-1}[(1, X, X^2, |X|, X^3, M_1, XM_1, X^2M_1, M_2, XM_2)\gamma_2] \\
\mathbb{E}[Y|X, A, M_1, M_2] &= (1, X, X^2, A, M_1, XM_1, M_2, AM_2)\theta_2; \quad \mu_2(X, M_1, M_2) = \mathbb{E}[Y|X, A = a, M_1, M_2] \\
\mathbb{E}[\mu_2(X, M_1, M_2)|X, A, M_1] &= (1, X, X^2, A, XA, M_1, XM_1)\theta_1; \quad \mu_1(X, M_1) = \mathbb{E}[\mu_2(X, M_1, M_2)|X, A = a_2, M_1] \\
\mathbb{E}[\mu_1(X, M_1)|X, A] &= (1, X, X^2, X^3, A, XA)\theta_0; \quad \mu_0(X) = \mathbb{E}[\mu_1(X, M_1)|X, A = a_1]
\end{aligned}$$

To demonstrate the multiple robustness of the EIF-based estimators, we also fit a misspecified model for each of the nuisance functions:

$$\pi_0(1|X) = \text{logit}^{-1}[(1, X)\tilde{\gamma}_0]$$

$$
\begin{aligned}
\pi_1(1|X, M_1) &= \text{logit}^{-1}[(1, X, M_1)\tilde{\gamma}_1] \\
\pi_2(1|X, M_1, M_2) &= \text{logit}^{-1}[(1, X, M_1, M_2)\tilde{\gamma}_2] \\
\mathbb{E}[Y|X, A, M_1, M_2] &= (1, X, A, M_1, M_2)\tilde{\theta}_2; \quad \mu_2(X, M_1, M_2) = \mathbb{E}[Y|X, A = a, M_1, M_2] \\
\mathbb{E}[\mu_2(X, M_1, M_2)|X, A, M_1] &= (1, X, A, M_1)\tilde{\theta}_1; \quad \mu_1(X, M_1) = \mathbb{E}[\mu_2(X, M_1, M_2)|X, A = a_2, M_1] \\
\mathbb{E}[\mu_1(X, M_1)|X, A] &= (1, X, A)\tilde{\theta}_0; \quad \mu_0(X) = \mathbb{E}[\mu_1(X, M_1)|X, A = a_1]
\end{aligned}
$$

Each of the five cases described in Section 5 reflects a combination of estimated nuisance functions from these correctly and incorrectly specified models. For example, in case (a), all parametric estimators of ψ_{010} use correctly specified models for $\pi_0(1|x), \pi_1(1|x, m_1), \pi_2(1|x, m_1, m_2)$ and incorrectly specified models for $\mu_0(x)$, $\mu_1(x, m_1)$, and $\mu_2(x, m_1, m_2)$. For the nonparametric estimators $\hat{\psi}_{\text{np}}^{\text{eif}_2}$ and $\hat{\psi}_{\text{tmle}}^{\text{eif}_2}$, all of the six nuisance functions are estimated via a super learner composed of Lasso and random forest, where the feature matrix consists of first-order, second-order, and interaction terms of the corresponding covariates. Two-fold cross-fitting is used to obtain the final estimates of ψ_{010}.