Semiparametric causal mediation analysis in cluster-randomized experiments

Chao Cheng[1,2,*] and Fan Li[1,2,†]

[1]Department of Biostatistics, Yale School of Public Health
[2]Center for Methods in Implementation and Prevention Science, Yale School of Public Health
[*]c.cheng@yale.edu [†]fan.f.li@yale.edu

April 30, 2024

Abstract

In cluster-randomized experiments, there is emerging interest in exploring the causal mechanism in which a cluster-level treatment affects the outcome through an intermediate outcome. Despite an extensive development of causal mediation methods in the past decade, only a few exceptions have been considered in assessing causal mediation in cluster-randomized studies, all of which depend on parametric model-based estimators. In this article, we develop the formal semiparametric efficiency theory to motivate several doubly-robust methods for addressing several mediation effect estimands corresponding to both the cluster-average and the individual-level treatment effects in cluster-randomized experiments—the natural indirect effect, natural direct effect, and spillover mediation effect. We derive the efficient influence function for each mediation effect, and carefully parameterize each efficient influence function to motivate practical strategies for operationalizing each estimator. We consider both parametric working models and data-adaptive machine learners to estimate the nuisance functions, and obtain semiparametric efficient causal mediation estimators in the latter case. Our methods are illustrated via extensive simulations and two completed cluster-randomized experiments.

Keywords: Double robustness; efficient influence function; Gaussian copula; natural indirect effect; machine learning; spillover mediation effect

1 Introduction

Cluster-randomized experiments are increasingly common in public health, social and educational research to study the population-level intervention effect, and each cluster, as the unit of randomization, could be a primary care clinic, school, or community. While the total (or average) treatment effect has been a cornerstone in the analysis of cluster-randomized experiments, there is an emerging interest in understanding the mechanisms by which a cluster-level treatment affects the outcome, and in elucidating the causal pathways for optimizing future interventions to enhance population benefits (Williams, 2016). Causal mediation analysis, with an aim to disentangle the role of a pre-specified mediator in explaining the observed treatment-outcome relationship, is well-suited for this objective. Although other definitions of causal mediation effects exist, a common approach is to decompose the total effect into a natural indirect effect through the mediator and a natural direct effect bypassing the mediator. Assessment of the natural (in)direct effects holds the promise to advance the theory underlying process evaluation, inform effective healthcare or social policy, and optimize future cluster-level interventions (Lee et al., 2021).

To assess mediation in cluster-randomized experiments, several methods have been developed to address the presence of within-cluster correlation and interference (Hudgens and Halloran, 2008). For example, assuming a continuous outcome and a continuous mediator, Bauer et al. (2006) and Zhang et al. (2009) considered the linear mixed models to estimate the natural indirect effect. Extending the work of Imai et al. (2010), Park and Kaplan (2015) provided a set of identification conditions for the natural indirect effect, and developed a Bayesian multilevel modeling approach for mediation analysis. These approaches implicitly invoke the no within-cluster interference assumption for the mediator, and impose strong parametric modeling assumptions on the data generating mechanism. Relaxing the no interference assumption for the mediator, VanderWeele (2010) and VanderWeele et al. (2013) provided a further decomposition of the natural indirect effect into a *spillover mediation effect* and an *individual mediation effect*, for which identification conditions and nonparametric identification formulas are provided. Each identification formula permits the use of multilevel models to derive the mediation effects, and the consistency of the final estimator critically depends on the correct specification of the fitted multilevel models. Relatedly, Forastiere et al. (2016) and Park and Kang (2023) have developed methods to address noncompliance in cluster-randomized experiments, where the treatment receipt can be viewed as a special binary mediator. However, their primary interest lies in the the network effects among different compliance subgroups, and addresses a different scientific question from the mediation context. Finally, all previous methods have assumed away *informative cluster size*, where the cluster size itself may be a surrogate of the within-cluster dynamics that is predictive of the mediator and/or the outcome and hence should be an intrinsic element of the estimands in cluster-randomized experiments (Kahan et al., 2023).

This article formalizes a semiparametric approach to assess causal mediation in cluster-randomized experiments that addresses the limitation of the previous developments. Our novel contributions are several-folded. First, we accommodate the general setup of informative cluster size, which gives rise to two versions of the total effect estimands. The *cluster-average treatment effect* targets the average change on the population of clusters along with their natural cluster members, whereas the *individual-average treatment effect* targets the average change on the population of all individuals across clusters. Depending on the scientific question and context, each estimand could be of interest and a comparative interpretation is discussed in Kahan et al. (2023). Based on the two total effect estimands, we define mediation effects in cluster-randomized experiment at both the cluster and individual level. Second, following VanderWeele et al. (2013), we allow for within-cluster interference and provide sufficient conditions to point identify the natural indirect effect, natural direct effect, as well as the spillover mediation effect that correspond to the decomposition of both the cluster-average and individual-average treatment effects. Leveraging the semiparametric efficiency theory (Bickel et al., 1993), we obtain the efficient influence function (EIF) of each estimand, and characterize the efficiency bound that is achievable by an optimal estimator. To improve the robustness of causal mediation analysis to model misspecification, we use the EIF to construct doubly robust estimators with parametric nuisance functions and efficient estimators when each nuisance is estimated via data-adaptive machine learners. Third, recognizing that the original EIF includes a multivariate mediator density, we propose two strategies to overcome the challenges from direct specifying a multivariate density model. The first strategy considers a copula representation of the joint mediator density and the second strategy reparameterizes the EIFs solely based on one-dimensional nuisance functions. Each strategy simplifies the specification of nuisance functions involved in the EIFs, leads to a doubly robust estimator for the natural (in)direct effect and a conditional doubly robust estimator for the spillover mediation effect, as well as admits efficient machine learning estimators under cross-fitting.

2 Notation and data structure

We consider a cluster-randomized experiment with K clusters. For cluster $i \in \{1, \ldots, K\}$, we define N_i as the number of individuals of that cluster (i.e., the cluster size), $A_i \in \{0, 1\}$ as the cluster-level treatment with 1 indicating the treated condition and 0 indicating the control condition, and $\boldsymbol{V}_i \in \mathbb{R}^{d_{\boldsymbol{V}} \times 1}$ as a vector of cluster-level baseline covariates. For individual $j \in \{1, \ldots, N_i\}$ of cluster i, we let $\boldsymbol{X}_{ij} \in \mathbb{R}^{d_{\boldsymbol{X}} \times 1}$ be a vector of individual-level baseline covariates, and write $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iN_i}]^T \in \mathbb{R}^{N_i \times d_{\boldsymbol{X}}}$. Let $\boldsymbol{C}_i = \{\boldsymbol{V}_i, \boldsymbol{X}_i\}$ be all baseline covariates in cluster i. We further define $Y_{ij} \in \mathbb{R}$ as the individual-level outcome measured at the end of study, $M_{ij} \in \mathbb{R}$ as the individual-level mediator that is measured before the outcome but after treatment assignment. We define $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iN_i}]^T \in \mathbb{R}^{N_i \times 1}$, $\boldsymbol{M}_i = [M_{i1}, \ldots, M_{iN_i}]^T \in \mathbb{R}^{N_i \times 1}$, and $\boldsymbol{M}_{i(-j)} \in \mathbb{R}^{(N_i - 1) \times 1}$ as a vector of mediators from cluster i excluding individual j. To summarize, we observe $\boldsymbol{O}_i = \{N_i, \boldsymbol{C}_i, A_i, \boldsymbol{M}_i, \boldsymbol{Y}_i\}$, $i = 1, \ldots, K$, where the causal relationship among these variables is illustrated in Figure 1(a). For conciseness, we sometimes omit the cluster indicator i in the subscript, such that $\boldsymbol{O}$ for $\boldsymbol{O}_i$, $Y_{\cdot j}$ for Y_{ij}, $M_{\cdot j}$ for M_{ij}, and $\boldsymbol{M}_{\cdot(-j)}$ for $\boldsymbol{M}_{i(-j)}$.

Under the potential outcomes framework, we define $M_{\cdot j}(a)$ as the potential mediator variable under condition $a \in \{0,1\}$, $\boldsymbol{M}(a) = [M_{\cdot 1}(a), \ldots, M_{\cdot N}(a)]^T$ as the vector of potential mediator variables, and $\boldsymbol{M}_{\cdot(-j)}(a)$ as the vector by excluding the jth element in $\boldsymbol{M}(a)$. We define $Y_{\cdot j}(a, \boldsymbol{m})$ as the potential outcome has the cluster been randomized to condition a and the mediators of all individuals in that cluster, $\boldsymbol{M}$, were set to value $\boldsymbol{m}$. Notice that one can equivalently represent $Y_{\cdot j}(a, \boldsymbol{m}) = Y_{\cdot j}(a, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)})$ with $\boldsymbol{m} = \{m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\}$; this notation explicitly distinguishes an individual's own mediator ($M_{\cdot j}$) from the mediators of the remaining cluster members ($\boldsymbol{M}_{\cdot(-j)}$). Furthermore, we assume composition (VanderWeele and Vansteelandt, 2009) such that $Y_{\cdot j}(a) = Y_{\cdot j}\left(a, \boldsymbol{M}(a)\right) = Y_{\cdot j}\left(a, M_{\cdot j}(a), \boldsymbol{M}_{\cdot(-j)}(a)\right)$; that is, the potential outcome under condition a is assumed identical to the potential outcome when the treatment condition is set to a and when all mediators in that cluster were set to their natural values under condition a.

In cluster-randomized experiments, clusters are frequently heterogeneous and cluster size can be associated with the potential outcome and mediators, leading to informative cluster size. For example, in health research, N may act as a proxy for provider volume and experience and thus can affect post-treatment variables. In social science research, N may correlate with participant demographics and socioeconomic status and can serve as effect modifiers for post-treatment variables. Without ruling out informative cluster, the total causal effect from the treatment to outcome can be measured through either the *cluster-average treatment effect* or *individual-average treatment effect* (Wang et al., 2023),

$$\mathrm{TE}_C = g\left(\mu_C(1), \mu_C(0)\right), \quad \mathrm{TE}_I = g\left(\mu_I(1), \mu_I(0)\right), \tag{1}$$

respectively, where $g(\cdot,\cdot)$ is a function determining the scale of effect measure, and

$$\mu_C(a) = \mathbb{E}\left\{\frac{\sum_{j=1}^N Y_{\cdot j}(a)}{N}\right\}, \quad \mu_I(a) = \frac{\mathbb{E}\left\{\sum_{j=1}^N Y_{\cdot j}(a)\right\}}{\mathbb{E}\{N\}}, \quad a \in \{0,1\}.$$

For example, $g(x,y) = x - y$, $g(x,y) = \frac{x}{y}$ and $g(x,y) = \frac{x/(1-x)}{y/(1-y)}$ correspond to causal mean difference, causal risk ratio and causal odds ratio, respectively. By construction,

$\mu_C(a)$ represents the average potential outcome under condition a for the population of all clusters, whereas $\mu_I(a)$ represents the average potential outcome under condition a among the population for all individuals across all clusters (Kahan et al., 2023); both can be of interest depending on the scientific question. Intuitively, TE_I resembles a natural estimand that one would have targeted under individual randomization, but TE_C is relatively more specialized to cluster randomization.

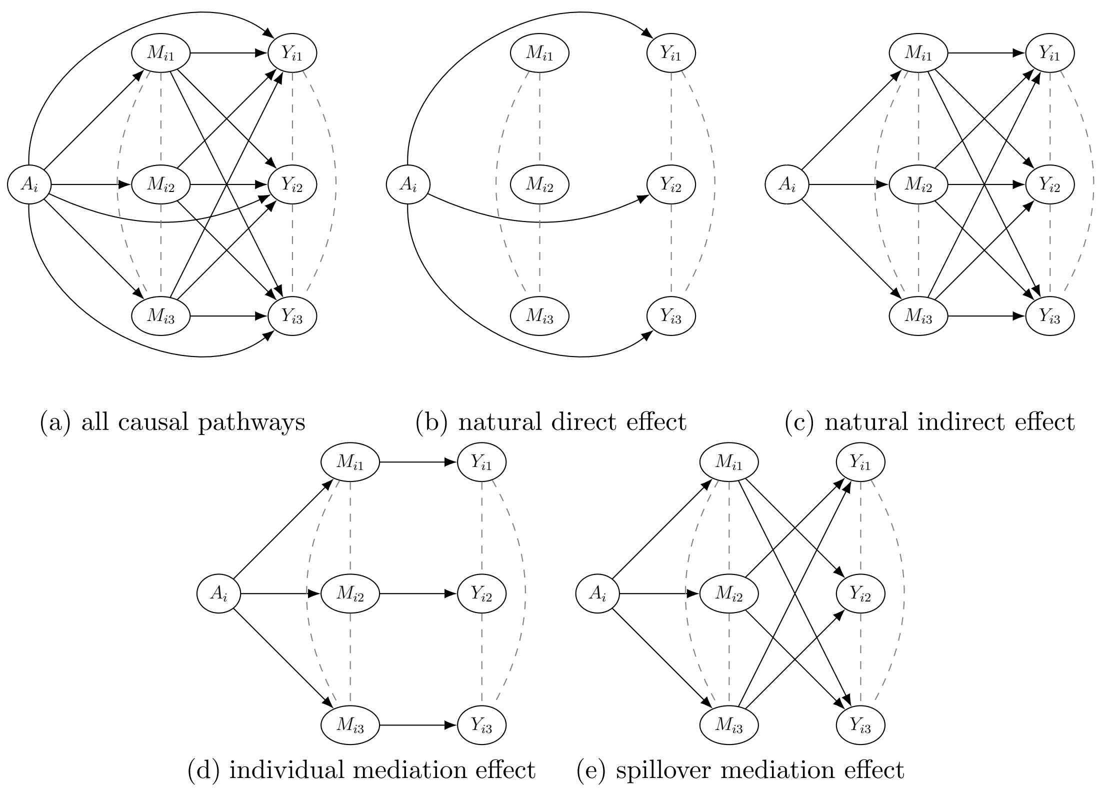

Figure 1: Directed acyclic graphs of the causal relationships among variables in a cluster with $N_i = 3$ individuals. A dashed edge indicates generic association with unknown causal structure. We omit all pre-treatment variables, $\{N_i, \boldsymbol{X}_i, \boldsymbol{V}_i\}$, and their associated causal pathways, but acknowledge that all pre-treatment variables should have direct pathways towards all mediators and outcomes (M_{ij} and Y_{ij} for all $j = 1, 2, 3$). Panel (a) includes all pathways from treatment to the outcome; Panel (b) collects all direct treatment pathways toward outcome (natural direct effect); Panel (c) collects all treatment pathways transmitted through the mediators (natural indirect effect); Panels (d) collects all treatment pathways toward outcome transmitted through each individual's own mediator (individual mediation effect); Panels (e) collects all treatment pathways toward outcome transmitted through mediators from all other same-cluster members (spillover mediation effect).

3 Causal estimands, assumptions, and identification

For ease of presentation, we focus on $g(x, y) = x - y$ so all estimands are defined on the difference scale, but extensions to ratio scales are straightforward and will be discussed in due course. We first decompose the cluster-average treatment effect into a *cluster-average natural indirect effect* (NIE_C) and a *cluster-average natural direct effect* (NDE_C):

$$\text{TE}_C = \underbrace{g\left(\theta_C(1,1), \theta_C(1,0)\right)}_{\text{NIE}_C} + \underbrace{g\left(\theta_C(1,0), \theta_C(0,0)\right)}_{\text{NDE}_C}, \tag{2}$$

with

$$\theta_C(a, a^*) = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(a, M_{\cdot j}(a^*), \boldsymbol{M}_{\cdot(-j)}(a^*)\right)\right] \text{ for } a, a^* \in \{0, 1\},$$

and $\theta_C(a, a) = \mu_C(a)$ by composition. NIE_C defines a contrast between cluster-average potential outcomes under treatment, by switching the mediators in that cluster from their counterfactual values under control to factual values under treatment. NDE_C compares the cluster-average potential outcomes under different treatment conditions, but fixing the potential mediators in that cluster to their values under control. Intuitively, NDE_C collects all direct causal pathways from treatment to outcome (Figure 1(b)) whereas NIE_C collects the remaining indirect causal pathways that must involve the mediator (Figure 1(c)).

The conventional decomposition (2) concerns $\boldsymbol{M}$ as an entirety for explaining the treatment-outcome mechanism. To disentangle the role of each individual's own mediator $M_{\cdot j}$ from that of other same-cluster members' mediator $\boldsymbol{M}_{\cdot(-j)}$, we follow VanderWeele et al. (2013) to further decompose NIE_C into a *cluster-average spillover mediation effect* (SME_C) and a *cluster-average individual mediation effect* (IME_C) as:

$$\text{NIE}_C = \underbrace{g\left(\theta_C(1,1), \tau_C\right)}_{\text{SME}_C} + \underbrace{g\left(\tau_C, \theta_C(1,0)\right)}_{\text{IME}_C}, \tag{3}$$

where $\tau_C = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(1, M_{\cdot j}(1), \boldsymbol{M}_{\cdot(-j)}(0)\right)\right]$. The IME_C estimand compares cluster-average potential outcomes under treatment, by switching the individual's mediator from its value under control to that under treatment, but maintaining the mediators from all remaining same-cluster members at their values without treatment. The SME_C estimand compares cluster-average potential outcomes under treatment, by instead switching the remaining same-cluster members' mediators from their values under control to those under treatment, but maintaining an individual's own mediator at its value under treatment. Thus, IME_C investigates the indirect effect explained by each individual's own mediator (illustrated by Figure 1(d)), whereas SME_C tackles the indirect or spillover effect explained by mediators of other individuals from the same cluster (illustrated by Figure 1(e)).

The above decomposition can be extended in two directions. First, the same decomposition applies to TE_I, as we can write $\text{TE}_I = \text{NIE}_I + \text{NDE}_I = g\left(\theta_I(1,1), \theta_I(1,0)\right) + g\left(\theta_I(1,0), \theta_I(0,0)\right)$ and $\text{NIE}_I = \text{SME}_I + \text{IME}_I = g\left(\theta_I(1,1), \tau_I\right) + g\left(\tau_I, \theta_I(1,0)\right)$. In this set of estimands, $\theta_I(a, a^*) = \mathbb{E}\left[\sum_{j=1}^N Y_{\cdot j}\left(a, M_{\cdot j}(a^*), \boldsymbol{M}_{\cdot(-j)}(a^*)\right)\right]/\mathbb{E}[N]$ for any $a, a^* \in \{0,1\}$ and $\tau_I = \mathbb{E}\left[\sum_{j=1}^N Y_{\cdot j}\left(1, M_{\cdot j}(1), \boldsymbol{M}_{\cdot(-j)}(0)\right)\right]/\mathbb{E}[N]$. Interpretations of individual-average causal mediation estimands $\{\text{NIE}_I, \text{NDE}_I, \text{IME}_I, \text{SME}_I\}$ are comparable to their cluster-average counterparts, but now correspond to a different total effect estimand. Second, the decomposition is easily generalized to ratio effect measures. For g chosen as either a risk ratio or odds ratio function, one can decompose TE_C in a multiplicative fashion such that $\text{TE}_C = \text{NIE}_C \times \text{NDE}_C$ and $\text{NIE}_C = \text{SME}_C \times \text{IME}_C$, where NIE_C, NDE_C, SME_C, and IME_C follow their counterparts in (2) and (3) except for setting $g(x,y) = \frac{x}{y}$ or $\frac{x/(1-x)}{y/(1-y)}$. Similar decomposition of TE_I can be carried out on the ratio scale and is omitted for brevity.

For ease of presentation, we focus on identification of $\theta_V(a, a^*)$ and τ_V for both $V \in \{C, I\}$, based on which all causal mediation effects are identified for any choice of g. We first state the requisite identification assumptions in cluster-randomized experiments.

Assumption 1. *(Consistency) $M_{ij}(a) = M_{ij}$ if $A_i = a$ and $Y_{ij}(a, m_{ij}, \boldsymbol{m}_{i(-j)}) = Y_{ij}$ if $A_i = a$ and $\{M_{ij}, \boldsymbol{M}_{i(-j)}\} = \{m_{ij}, \boldsymbol{m}_{i(-j)}\}$, for all i, j, $a \in \{0,1\}$ and $\{m_{ij}, \boldsymbol{m}_{i(-j)}\}$ over their valid support.*

Assumption 2. *(Cluster randomization) Treatment A is randomized at the cluster level such that A_i is independently drawn from a Bernoulli trial with $\mathbb{P}(A_i = 1) = \pi \in (0, 1)$.*

Assumption 3. *(Sequential ignorability) $\{\boldsymbol{M}_i(1), \boldsymbol{M}_i(0)\} \perp Y_{ij}(a, m_{ij}, \boldsymbol{m}_{i(-j)})|\{A_i, \boldsymbol{C}_i, N_i\}$, for all i, j, $a \in \{0, 1\}$ and $\{m_{ij}, \boldsymbol{m}_{i(-j)}\}$ over their valid support.*

Assumption 4. *(Super-population) $\boldsymbol{O}_1,\ldots,\boldsymbol{O}_K$ are mutually independent. For each cluster, the cluster size N follows a distribution P_N over a finite support on $\mathbb{N}^+$. Conditional on N, the joint distribution $P_{\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C}|N}$ can be decomposed as $P_{\boldsymbol{Y}|A,\boldsymbol{M},\boldsymbol{C},N} \times P_{\boldsymbol{M}|A,\boldsymbol{C},N} \times P_A \times P_{\boldsymbol{C}|N}$ with each component having a finite second moment. Furthermore, positivity holds such that $f_{\boldsymbol{M}|A,\boldsymbol{C},N}(\boldsymbol{m}|a, \boldsymbol{c}, n) > 0$ for any $\{\boldsymbol{m}, a, \boldsymbol{c}, n\}$ over their valid support.*

Assumption 1 connects the potential values of the outcome and mediator to their observed counterparts, allows for mediator interference within each cluster but rules out interference across clusters. Assumption 2 requires $A_i \perp \{N_i, \boldsymbol{C}_i, \boldsymbol{M}_i(1), \boldsymbol{M}_i(0), Y_{ij}(a, m_{ij}, \boldsymbol{m}_{i(-j)})\}$ to eliminate unmeasured confounding for both the treatment-mediator and the treatment-outcome relationships, which holds by the cluster-randomization study design. Assumption 3 extends the sequential ignorability assumption in standard causal mediation analysis with independent data (Imai et al., 2010) to clustered data, and assumes away unmeasured confounding for the mediator-outcome relationship. Similar to Wang et al. (2023), Assumption 4 conceptualizes a super-population of clusters, by which $\boldsymbol{O}_1, \ldots, \boldsymbol{O}_K$, are independent drawn from the mixture distribution $P_{\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C}|N} \times P_N$, where the support of cluster size N is assumed to be bounded. Assumptions 1–4 are sufficient for identifying $\theta_V(a, a^*)$ for $V \in \{C, I\}$. But to point identify τ_V, the following additional assumption is required.

Assumption 5. *(Inter-individual cross-world mediator independence) Conditional on cluster size and all baseline covariates, $M_{ij}(1) \perp \boldsymbol{M}_{i(-j)}(0)|\{\boldsymbol{C}_i, N_i\}$ for all i and j.*

Assumption 5 requires that, after adjusting for $\boldsymbol{C}_i$ and N_i, an individual's potential mediator under treatment is independent from the potential mediators from all other individuals in the same cluster under control. Importantly, given $\boldsymbol{C}_i$ and N_i, Assumption 5 still allows for arbitrary residual dependence between single-world potential mediators within the same cluster (i.e., $M_{ij}(a)$ and $M_{ij'}(a)$ can be arbitrarily correlated) and intra-individual cross-world mediator dependence (i.e., $M_{ij}(a)$ and $M_{ij}(a^*)$ with $a \neq a^*$ can be arbitrarily correlated), which are more likely to be present in cluster-randomized experiments.

To proceed with nonparametric identification, we introduce four nuisance functions of $\boldsymbol{O}$. Specifically, define $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n) = f_{\boldsymbol{M}|A,\boldsymbol{C},N}(\boldsymbol{m}|a, \boldsymbol{c}, n)$ as the joint density (probability) of mediators in a cluster conditional on treatment assignment and baseline information. Let $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n) = f_{M_{\cdot j}|A,\boldsymbol{C},N}(m_{\cdot j}|a, \boldsymbol{c}, n)$ and $\kappa_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = f_{\boldsymbol{M}_{\cdot(-j)}|A,\boldsymbol{C},N}(\boldsymbol{m}_{\cdot(-j)}|a, \boldsymbol{c}, n)$ be the corresponding densities of $M_{\cdot j}$ and $\boldsymbol{M}_{\cdot(-j)}$. Notice that $\kappa_{\cdot j}$ and $\kappa_{\cdot(-j)}$ can be derived from a joint mediator density with $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n) = \int_{\boldsymbol{m}_{\cdot(-j)}} \kappa(a, \boldsymbol{m}, \boldsymbol{c}, n) \mathrm{d}\boldsymbol{m}_{\cdot(-j)}$ and $\kappa_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = \int_{m_{\cdot j}} \kappa(a, \boldsymbol{m}, \boldsymbol{c}, n) \mathrm{d}m_{\cdot j}$. Define $\eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \mathbb{E}[Y_{\cdot j}|A = a, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{C} = \boldsymbol{c}, N = n]$ as the expectation of $Y_{\cdot j}$ conditional on treatment assignment, mediator, and baseline information in that cluster. We abbreviate the nuisance functions as $h^{(1)}_{nuisance} = \{\eta_{\cdot j}, \kappa, \kappa_{\cdot j}, \kappa_{\cdot(-j)}\}$. For a quick reference, the list of all nuisance functions required are summarized in Table 1. These definitions then prepare us to introduce the following nonparametric identification formulas.

Theorem 1. *(Nonparametric identification) Under Assumptions 1–4, we can identify*

$$\theta_C(a, a^*) = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N)\kappa(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\, d\boldsymbol{m}\right],$$

$$\theta_I(a, a^*) = \mathbb{E}\left[\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N)\kappa(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\, d\boldsymbol{m}\right] \Big/ \mathbb{E}[N],$$

for any $a, a^* \in \{0, 1\}$. Additionally, if Assumption 5 holds, τ_C and τ_I can be identified by

$$\tau_C = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\kappa_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\, d\boldsymbol{m}\right],$$

$$\tau_I = \mathbb{E}\left[\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\kappa_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\, d\boldsymbol{m}\right] \Big/ \mathbb{E}[N].$$

Theorem 1 generalizes the identification formulas in VanderWeele et al. (2013) to the scenario with informative cluster size, and explicitly distinguishes cluster-level and individual-level inference to address different levels of hypothesis Kahan et al. (2023); Wang et al. (2023). Based on Theorem 1, all mediation effects are identified as they are known functions of $\theta_V(a, a^*)$ and τ_V ($V \in \{C, I\}$). Following Tchetgen and Shpitser (2012), we shall refer to the nonparametric identification formulas in Theorem 1 as *mediation functionals*.

Remark 1. *(Interventional effect interpretation) When Assumption 5 fails to hold, we show in the Supplementary Material that each mediation functional remains to be causally interpretable under the interventional causal mediation framework (Moreno-Betancur and Carlin, 2018). For example, under Assumptions 1–4 only,* $\theta_C(a, a^*) = \theta_C^{(int)}(a, a^*) := \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(a, G_{\cdot j}(a^*), \boldsymbol{G}_{\cdot(-j)}(a^*)\right)\right]$ *and* $\tau_C = \tau_C^{(int)} := \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(1, G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(0)\right)\right]$, *where* $\boldsymbol{G}(1) = \{G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(1)\}$ *and* $\boldsymbol{G}(0) = \{G_{\cdot j}(0), \boldsymbol{G}_{\cdot(-j)}(0)\}$ *are two independent random draws of* $\boldsymbol{M}(1)$ *and* $\boldsymbol{M}(0)$ *from* $P_{\boldsymbol{M}(1)|\boldsymbol{C},N}$ *and* $P_{\boldsymbol{M}(0)|\boldsymbol{C},N}$*, respectively. Then* $SME_C = g(\theta_C^{(int)}(1,1), \tau_C^{(int)})$ *is interpreted as the cluster-average interventional spillover mediation effect (*$ISME_C$*), which measures the change in cluster-average potential outcomes if the distribution of potential mediators from all other same-cluster members were shifted from what they would have been under control to those under treatment, while fixing every individual's own potential mediator distribution to what it would have been under treatment. Similarly,* $IME_C = g(\tau_C^{(int)}, \theta_C^{(int)}(1,0))$ *is the cluster-average interventional individual mediation effect (*$IIME_C$*) which can be interpreted in a similar fashion.*

4 Semiparametric and nonparametric estimation

4.1 Specification of parametric working models

We first consider parametric models to estimate $h_{nuisance}^{(1)}$. Since $\{\kappa_{\cdot j}, \kappa_{\cdot(-j)}\}$ can be specified from κ, only $\{\eta_{\cdot j}, \kappa\}$ need to be modeled. For $\eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)$, one can specify a mean model of Y_{ij} conditional on A_i, $\boldsymbol{M}_i$, $\boldsymbol{C}_i = \{\boldsymbol{X}_i, \boldsymbol{V}_i\}$, and N_i. Given that the dimensions

Table 1: A summary of definitions of nuisance functions and their requirement in constructing estimators based on the efficient influence functions.¶‖

Notation	Mathematical definition	$\widehat{\theta}_V^{\text{eif}_1}(a,a^*)$	$\widehat{\theta}_V^{\text{eif}_2}(a,a^*)$	$\widehat{\tau}_V^{\text{eif}_1}$	$\widehat{\tau}_V^{\text{eif}_2}$
$\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$	$\mathbb{E}[Y_{\cdot j}\|A=a,\boldsymbol{M}=\boldsymbol{m},\boldsymbol{C}=\boldsymbol{c},N=n]$	✓	✓	✓	✓
$\eta^{\star}_{\cdot j}(a,a^*,\boldsymbol{c},n)$	$\mathbb{E}\left[\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\|A=a^*,\boldsymbol{C}=\boldsymbol{c},N=n\right]$		✓		
$\eta^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$	$\mathbb{E}[\delta(\boldsymbol{M},\boldsymbol{C},N)\|A=a^*,M_{\cdot j}=m_{\cdot j},\boldsymbol{C}=\boldsymbol{c},N=n]$§				✓
$\kappa(a,\boldsymbol{m},\boldsymbol{c},n)$	$f_{\boldsymbol{M}\|A,\boldsymbol{C},N}(\boldsymbol{m}\|a,\boldsymbol{c},n)$	✓		✓	
$\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)$	$f_{M_{\cdot j}\|A,\boldsymbol{C},N}(m_{\cdot j}\|a,\boldsymbol{c},n)$			✓	✓
$\kappa^{\star}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$	$f_{M_{\cdot j}\|\boldsymbol{M}_{\cdot(-j)},A,\boldsymbol{C},N}(m_{\cdot j}\|\boldsymbol{m}_{\cdot(-j)},a,\boldsymbol{c},n)$				✓
$\kappa_{\cdot(-j)}(a,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)$	$f_{\boldsymbol{M}_{\cdot(-j)}\|A,\boldsymbol{C},N}(\boldsymbol{m}_{\cdot j}\|a,\boldsymbol{c},n)$			✓	
$s(a,\boldsymbol{m},\boldsymbol{c},n)$	$f_{A\|\boldsymbol{M},\boldsymbol{C},N}(a\|\boldsymbol{m},\boldsymbol{c},n)$		✓		✓

¶ For notation brevity, the superscript '-par' or '-ml' in estimators are omitted.

‖ Nuisance functions used in $\widehat{\theta}_V^{\text{mf-par}}(a,a^*)$ and $\widehat{\tau}_V^{\text{mf-par}}$ are identical to these used in $\widehat{\theta}_V^{\text{eif}_1}(a,a^*)$ and $\widehat{\tau}_V^{\text{eif}_1}$.

§ The function is defined as $\delta(\boldsymbol{M},\boldsymbol{C},N):=\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\kappa_{\cdot j}(a,M_{\cdot j},\boldsymbol{C},N)/\kappa^{\star}_{\cdot j}(a^*,\boldsymbol{M},\boldsymbol{C},N)$.

of $\boldsymbol{M}_i$ and $\boldsymbol{X}_i$ can vary across clusters, a practical strategy is to adjust for summary functions with fixed dimensions in the regression model (Ogburn et al., 2022). For example, a bivariate summary function $\left\{M_{ij},\frac{1}{N_i-1}\sum_{l=1,l\neq j}^{N_i}M_{il}\right\}$ of $\boldsymbol{M}_i$ can be considered such that Y_{ij} is assumed to be affected by $\boldsymbol{M}_i$ via one's own mediator and the average mediator values of other same-cluster members. Similarly, $\left\{\boldsymbol{X}_{ij},\frac{1}{N_i-1}\sum_{l=1,l\neq j}^{N_i}\boldsymbol{X}_{il}\right\}$ can be used for $\boldsymbol{X}_i$.

Modeling $\kappa(a,\boldsymbol{m},\boldsymbol{c},n)$ requires specification of a joint density of mediators among all same-cluster individuals. We write out $\kappa(a,\boldsymbol{m},\boldsymbol{c},n)$ with two variationally independent components: the marginal mediator probability for each individual $\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)$ and a copula $\mathcal{C}$ characterizing the association structure across $\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)$ for $j=1,\ldots,n$. By the Sklar's theorem (Joe, 1997), there exists a copula $\mathcal{C}$ such that

$$P_{\boldsymbol{M}|A,\boldsymbol{C},N}(\boldsymbol{m}|a,\boldsymbol{c},n)=\mathcal{C}(\mathcal{K}_{\cdot 1}(m_{\cdot 1}),\cdots,\mathcal{K}_{\cdot n}(m_{\cdot n})|a,\boldsymbol{c},n), \tag{4}$$

where $\mathcal{K}_{\cdot j}(m_{\cdot j}):=\int_{-\infty}^{m_{\cdot j}}\kappa_{\cdot j}(a,t,\boldsymbol{c},n)\mathrm{d}t$ is the CDF of $M_{\cdot j}$ and $\mathcal{C}(u_{\cdot 1},\ldots,u_{\cdot n}|a,\boldsymbol{c},n)$ is a n-variate copula supported on $\{u_{\cdot 1},\cdots,u_{\cdot n}\}\in[0,1]^n$. When the mediator is continuous, we have $\kappa(a,\boldsymbol{m},\boldsymbol{c},n)=\kappa^c(a,\boldsymbol{m},\boldsymbol{c},n):=c(\mathcal{K}_{\cdot 1}(m_{\cdot 1}),\cdots,\mathcal{K}_{\cdot n}(m_{\cdot n})|a,\boldsymbol{c},n)\prod_{j=1}^{n}\kappa_j(a,m_{\cdot j},\boldsymbol{c},n)$ and $\kappa_{\cdot(-j)}(a,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)=\kappa^c_{\cdot(-j)}(a,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n):=\int\kappa^c(a,\boldsymbol{m},\boldsymbol{c},n)\mathrm{d}m_{\cdot j}$ respectively, where $c(u_{\cdot 1},\ldots,u_{\cdot n}|a,\boldsymbol{c},n)=\frac{\partial^n}{\partial u_{\cdot 1}\cdots\partial u_{\cdot n}}\mathcal{C}(u_{\cdot 1},\ldots,u_{\cdot n}|a,\boldsymbol{c},n)$ is the density of the copula $\mathcal{C}$. When the mediator is discrete, expressions of $\kappa^c(a,\boldsymbol{m},\boldsymbol{c},n)$ and $\kappa^c_{\cdot(-j)}(a,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)$ necessitate multi-dimensional integrals with details provided in the Supplementary Material.

To proceed, one can specify a working marginal regression model to obtain $\widehat{\kappa}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$. For example, a linear model and a generalized linear model can be used to specify $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ when the mediator is continuous and discrete. The unknown parameters associated with $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ can then be estimated with the generalized estimating equations and a working independence assumption. In addition, a parametric multivariate copula model can be considered to specify the association structure. We focus on a Gaussian copula (Masarotto and Varin, 2012) by assuming

$$\mathcal{C}(u_{\cdot 1}, \cdots, u_{\cdot n}|a, \boldsymbol{c}, n) = \boldsymbol{\Phi}_n\Big([\Phi^{-1}(u_{\cdot 1}), \ldots, \Phi^{-1}(u_{\cdot n})]; \boldsymbol{I}_n(\boldsymbol{\rho})\Big), \tag{5}$$

where $\boldsymbol{I}_n(\boldsymbol{\rho}) \equiv \boldsymbol{I}_n(a, \boldsymbol{c}, n; \boldsymbol{\rho})$ is a n-by-n correlation matrix conditional on $A = a, \boldsymbol{C} = \boldsymbol{c}, N = n$ with a vector of unknown association parameters $\boldsymbol{\rho}$, $\phi(w)$ and $\Phi(w)$ are the density function and CDF for a univariate standard normal variable, $\boldsymbol{\phi}_n(\boldsymbol{w}; \boldsymbol{P})$ and $\boldsymbol{\Phi}_n(\boldsymbol{w}; \boldsymbol{P})$ are their analogues for a n-dimensional multivariate standard normal variable with correlation matrix $\boldsymbol{P}$. A common choice of $\boldsymbol{I}_n(\boldsymbol{\rho})$ is the exchangeable structure with a single correlation parameter $\rho \in (-1, 1)$ characterizing each off-diagonal element. Under the Gaussian copula and given $\widehat{\kappa}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$, $\boldsymbol{\rho}$ can be estimated by the pseudo-likelihood approach outlined in Supplementary Material; estimates of $\{\kappa, \kappa_{\cdot(-j)}\}$ are then obtained accordingly. Specific expressions for $\widehat{\kappa}^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ and $\widehat{\kappa}^c_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n)$ are presented in the Supplementary Material. Finally, in the special case of a continuous mediator modeled by linear regression and an exchangeable association structure, the induced parameterization for $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n)$ is precisely a linear mixed model and ρ is referred to as the intracluster correlation parameter (Murray et al., 1998). To summarize, the nuisance functions $h^{(1)}_{nuisance} = \{\eta_{\cdot j}, \kappa, \kappa_{\cdot j}, \kappa_{\cdot(-j)}\}$ can be estimated by $\widehat{h}^{(1)}_{nuisance} = \{\widehat{\eta}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^c_{\cdot(-j)}\}$ based on three parametric working models of $\eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)$, $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$, and $\mathcal{C}(u_{\cdot 1}, \ldots, u_{\cdot n}|a, \boldsymbol{c}, n)$.

The mediation functionals in Theorem 1 provide a way to estimate $\theta_V(a, a^*)$ and τ_V by replacing $h^{(1)}_{nuisance}$ with $\widehat{h}^{(1)}_{nuisance}$, and then substituting the expectation operator by the empirical cluster-average operator $\mathbb{P}_K[V] = \frac{1}{K}\sum_{i=1}^K V_i$. Specifically, $\theta_C(a,a^*)$ and τ_C are estimated by $\widehat{\theta}_C^{\text{mf-par}}(a,a^*) = \frac{1}{K}\sum_{i=1}^K \frac{1}{N_i}\sum_{j=1}^{N_i} \mathcal{I}_{1,ij}$ and $\widehat{\tau}_C^{\text{mf-par}} = \frac{1}{K}\sum_{i=1}^K \frac{1}{N_i}\sum_{j=1}^{N_i} \mathcal{I}_{2,ij}$, where

$$\mathcal{I}_{1,ij} = \int_{\boldsymbol{m}_i} \widehat{\eta}_{ij}(a,\boldsymbol{m}_i,\boldsymbol{C}_i,N_i)\widehat{\kappa}^c(a^*,\boldsymbol{m}_i,\boldsymbol{C}_i,N_i)\mathrm{d}\boldsymbol{m}_i,$$
$$\mathcal{I}_{2,ij} = \int_{\boldsymbol{m}_i} \widehat{\eta}_{ij}(1,\boldsymbol{m}_i,\boldsymbol{C}_i,N_i)\widehat{\kappa}_{ij}(1,m_{ij},\boldsymbol{C}_i,N_i)\widehat{\kappa}^c_{i(-j)}(0,\boldsymbol{m}_{i(-j)},\boldsymbol{C}_i,N_i)\mathrm{d}\boldsymbol{m}_i.$$

Analogously, we have $\widehat{\theta}_I^{\text{mf-par}}(a,a^*) = \frac{1}{K\times\overline{N}}\sum_{i=1}^K\sum_{j=1}^{N_i}\mathcal{I}_{1,ij}$ and $\widehat{\tau}_I^{\text{mf-par}} = \frac{1}{K\times\overline{N}}\sum_{i=1}^K\sum_{j=1}^{N_i}\mathcal{I}_{2,ij}$ for $\theta_I(a,a^*)$ and τ_C, where $\overline{N} = \frac{1}{K}\sum_{i=1}^K N_i$ is the average cluster size. The integrals, $\mathcal{I}_{1,ij}$ and $\mathcal{I}_{2,ij}$, can be calculated through Monte Carlo integration as follows. Let $\widetilde{\boldsymbol{m}}_i^{(r)}(1)$, for $r \in \{1,\ldots,R\}$, be R (e.g., $R = 100$) independent draws of $\boldsymbol{M}_i$ from $\widehat{\kappa}^c(1,\boldsymbol{m}_i,\boldsymbol{C}_i,N_i)$. Similarly, let $\widetilde{\boldsymbol{m}}_i^{(r)}(0)$, for $r \in \{1,\ldots,R\}$, be R independent draws of $\boldsymbol{M}_i$ from $\widehat{\kappa}^c(0,\boldsymbol{m}_i,\boldsymbol{C}_i,N_i)$. Then, $\mathcal{I}_{1,ij} \approx \frac{1}{R}\sum_{r=1}^R \widehat{\eta}_{ij}(a,\widetilde{\boldsymbol{m}}_i^{(r)}(a^*),\boldsymbol{C}_i,N_i)$ and $\mathcal{I}_{2,ij} \approx \frac{1}{R}\sum_{r=1}^R \widehat{\eta}_{ij}(1,\widetilde{m}_{ij}^{(r)}(1),\widetilde{\boldsymbol{m}}_{i(-j)}^{(r)}(0),\boldsymbol{C}_i,N_i)$, $\widetilde{\boldsymbol{m}}_{i(-j)}^{(r)}(0)$ is obtained from $\widetilde{\boldsymbol{m}}_i^{(r)}(0)$ by excluding its j-th element.

4.2 Semiparametric doubly robust estimators

To improve upon the fully parametric estimators via the mediation functionals, we leverage the theory of semiparametric inference (Bickel et al., 1993) to develop more robust estimators of each causal mediation estimand. To proceed, we first derive the efficient influence function (EIF) of each mediation estimand under a nonparametric model, where the observed data likelihood is left unrestricted (except for the known treatment assignment). The variance of the EIF then characterizes the semiparametric efficiency lower bound, i.e., the variance bound among any regular, consistent and asymptotically linear estimator. To introduce the EIFs, we first define the following six auxiliary functions of $h^{(1)}_{nuisance}$:

$$w^{(1)}(a,a^*,\boldsymbol{m},\boldsymbol{c},n) = \frac{\kappa(a^*,\boldsymbol{m},\boldsymbol{c},n)}{\kappa(a,\boldsymbol{m},\boldsymbol{c},n)}, \quad w^{(2)}(a,a^*,a',\boldsymbol{m},\boldsymbol{c},n) = \frac{\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)\kappa_{\cdot(-j)}(a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)}{\kappa(a',\boldsymbol{m},\boldsymbol{c},n)},$$
$$u^{(1)}_{\cdot j}(a,a^*,\boldsymbol{c},n) = \int_{\boldsymbol{m}} \eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)\kappa(a^*,\boldsymbol{m},\boldsymbol{c},n)\mathrm{d}\boldsymbol{m},$$

$$u_{\cdot j}^{(2)}(a,a^*,m_j,\boldsymbol{c},n)=\int_{\boldsymbol{m}_{\cdot(-j)}}\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)\kappa_{\cdot(-j)}(a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}_{\cdot(-j)}, \tag{6}$$

$$u_{\cdot j}^{(3)}(a,a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)=\int_{m_{\cdot j}}\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)\kappa_{\cdot j}(a^*,m_j,\boldsymbol{c},n)\mathrm{d}m_{\cdot j},$$

$$u_{\cdot j}^{(4)}(a,a^*,a',\boldsymbol{c},n)=\int_{\boldsymbol{m}}\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)\kappa_{\cdot j}(a^*,m_{\cdot j},\boldsymbol{c},n)\kappa_{\cdot(-j)}(a',\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)\mathrm{d}\boldsymbol{m},$$

where $\{w^{(1)},w^{(2)}\}$ are ratios of mediator densities and $\{u_{\cdot j}^{(1)},u_{\cdot j}^{(2)},u_{\cdot j}^{(3)},u_{\cdot j}^{(4)}\}$ are integrals of $\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$ with respect to distribution of the mediators under different assignments.

Theorem 2. *Under Assumptions 1–4, the EIFs of* $\theta_C(a,a^*)$ *and* $\theta_I(a,a^*)$ *are* $\mathcal{D}_{\theta_C(a,a^*)}(\boldsymbol{O})=\psi_\theta(a,a^*;\boldsymbol{O})-\theta_C(a,a^*)$ *and* $\mathcal{D}_{\theta_I(a,a^*)}(\boldsymbol{O})=\frac{N}{\mathbb{E}[N]}\{\psi_\theta(a,a^*;\boldsymbol{O})-\theta_I(a,a^*)\}$*, respectively, where*

$$\begin{aligned}\psi_\theta(a,a^*;\boldsymbol{O})=&\frac{1}{N}\sum_{j=1}^N\Bigg\{\frac{\mathbb{I}(A=a)}{\pi^a(1-\pi)^{1-a}}w^{(1)}(a,a^*,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j}-\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\\&+\frac{\mathbb{I}(A=a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left\{\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)-u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N)\right\}+u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N)\Bigg\}.\end{aligned}$$

Additionally, if Assumption 5 holds, the EIFs of τ_C *and* τ_I *are* $\mathcal{D}_{\tau_C}(\boldsymbol{O})=\psi_\tau(\boldsymbol{O})-\tau_C$ *and* $\mathcal{D}_{\tau_I}(\boldsymbol{O})=\frac{N}{\mathbb{E}[N]}\{\psi_\tau(\boldsymbol{O})-\tau_I\}$*, respectively, where*

$$\begin{aligned}\psi_\tau(\boldsymbol{O})=&\frac{1}{N}\sum_{j=1}^N\Bigg\{\frac{A}{\pi}w^{(2)}(1,0,1,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j}-\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\\&+\frac{A}{\pi}\left\{u_{\cdot j}^{(2)}(1,0,M_{\cdot j},\boldsymbol{C},N)-u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}\\&+\frac{1-A}{1-\pi}\left\{u_{\cdot j}^{(3)}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)-u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}+u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\Bigg\}.\end{aligned}$$

Therefore, the semiparametric efficiency lower bound for each causal estimand is $\mathbb{E}[\{\mathcal{D}_\zeta(\boldsymbol{O})\}^2]$*, for* $\zeta\in\{\theta_C(a,a^*),\theta_I(a,a^*),\tau_C,\tau_I\}$.

Theorem 2 shows that the EIFs for the individual-average mediation functionals bear a similar pattern to their cluster-average counterparts, except that the former additionally include a cluster size multiplier, $N/\mathbb{E}[N]$, to target the population of all individual units. Importantly, the EIFs are functions of $h_{nuisance}^{(1)}$ through the six auxiliary functions $\{w^{(1)},w^{(2)},u_{\cdot j}^{(1)},u_{\cdot j}^{(2)},u_{\cdot j}^{(3)},u_{\cdot j}^{(4)}\}$ defined in (6), and therefore directly motivate new causal mediation estimators that optimally combine information across all working models. Based on the working models in Section 4.1, we propose the following semiparametric estimators:

$$\widehat{\theta}_C^{\text{eif}_1\text{-par}}(a,a^*) = \frac{1}{K}\sum_{i=1}^{K}\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i), \quad \widehat{\tau}_C^{\text{eif}_1\text{-par}} = \frac{1}{K}\sum_{i=1}^{K}\widehat{\psi}_\tau(\boldsymbol{O}_i),$$

$$\widehat{\theta}_I^{\text{eif}_1\text{-par}}(a,a^*) = \frac{1}{K\times\overline{N}}\sum_{i=1}^{K}N_i\times\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i), \quad \widehat{\tau}_I^{\text{eif}_1\text{-par}} = \frac{1}{K\times\overline{N}}\sum_{i=1}^{K}N_i\times\widehat{\psi}_\tau(\boldsymbol{O}_i), \quad (7)$$

where $\{\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i),\widehat{\psi}_\tau(\boldsymbol{O}_i)\}$ are plugin estimators of $\{\psi_\theta(a,a^*;\boldsymbol{O}_i),\psi_\tau(\boldsymbol{O}_i)\}$ based on $\widehat{h}^{(1)}_{nuisance}$. Evaluation of $\{\widehat{u}^{(1)}_{\cdot j},\widehat{u}^{(2)}_{\cdot j},\widehat{u}^{(3)}_{\cdot j},\widehat{u}^{(2)}_{\cdot j}\}$ in $\{\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i),\widehat{\psi}_\tau(\boldsymbol{O}_i)\}$ require calculating multivariate integrals, and is obtained from Monte Carlo integration as in Section 4.1.

While the semiparametric estimators (7) directly follow from the EIF, they require the specification of a joint model for the within-cluster mediators $\kappa(a,\boldsymbol{m},\boldsymbol{c},n)$, and further necessitate the calculation of the multi-dimensional integrals with respect to the mediator density functions. To further address these two potential challenges, Proposition 1 provides an alternative strategy to carefully reparameterize the six auxiliary functions in (6).

Proposition 1. *(Low-dimensional reparameterization) Define*

$$s(a,\boldsymbol{m},\boldsymbol{c},n) = f_{A|\boldsymbol{M},\boldsymbol{C},N}(a|\boldsymbol{m},\boldsymbol{c},n), \quad \kappa^\star_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n) = f_{M_{\cdot j}|\boldsymbol{M}_{\cdot(-j)},A,\boldsymbol{C},N}(m_{\cdot j}|\boldsymbol{m}_{\cdot(-j)},a,\boldsymbol{c},n),$$

$$\eta^\star_{\cdot j}(a,a^*,\boldsymbol{c},n) = \mathbb{E}\left[\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)|A=a^*,\boldsymbol{C}=\boldsymbol{c},N=n\right],$$

$$\eta^\dagger_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n) = \mathbb{E}\left[\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\frac{\kappa_{\cdot j}(a,M_{\cdot j},\boldsymbol{C},N)}{\kappa^\star_{\cdot j}(a^*,\boldsymbol{M},\boldsymbol{C},N)}\Big|A=a^*,M_{\cdot j}=m_{\cdot j},\boldsymbol{C}=\boldsymbol{c},N=n\right].$$

Then, the auxiliary functions in (6) *can be re-expressed as:*

$$w^{(1)}(a,a^*,\boldsymbol{m},\boldsymbol{c},n) = \frac{s(a^*,\boldsymbol{m},\boldsymbol{c},n)}{s(a,\boldsymbol{m},\boldsymbol{c},n)}\times\frac{\pi^a(1-\pi)^{1-a}}{\pi^{a^*}(1-\pi)^{1-a^*}},$$

$$w^{(2)}(a,a^*,a',\boldsymbol{m},\boldsymbol{c},n) = \frac{\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)}{\kappa^\star_{\cdot j}(a^*,\boldsymbol{m},\boldsymbol{c},n)}\times\frac{s(a^*,\boldsymbol{m},\boldsymbol{c},n)}{s(a',\boldsymbol{m},\boldsymbol{c},n)}\times\frac{\pi^{a'}(1-\pi)^{1-a'}}{\pi^{a^*}(1-\pi)^{1-a^*}},$$

$u^{(1)}_{\cdot j}(a,a^*,\boldsymbol{c},n) = \eta^\star_{\cdot j}(a,a^*,\boldsymbol{c},n)$, $u^{(2)}_{\cdot j}(a,a^*,m_j,\boldsymbol{c},n) = \eta^\dagger_{\cdot j}(a,a^*,m_j,\boldsymbol{c},n)$, $u^{(3)}_{\cdot j}(a,a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)$ *remains unchanged, and* $u^{(4)}_{\cdot j}(a,a^*,a',\boldsymbol{c},n) = \int_{m_{\cdot j}}\eta^\dagger_{\cdot j}(a,a',m_j,\boldsymbol{c},n)\kappa_{\cdot j}(a^*,m_{\cdot j},\boldsymbol{c},n)dm_{\cdot j}$.

Proposition 1 permits us to consider an alternative set of nuisance functions $h^{(2)}_{nuisance}=$

$\{\eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n), \eta^{\star}_{\cdot j}(a, a^*, \boldsymbol{c}, n), \eta^{\dagger}_{\cdot j}(a, a^*, m_{\cdot j}, \boldsymbol{c}, n), \kappa_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n), \kappa^{\star}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n), s(a, \boldsymbol{m}, \boldsymbol{c}, n)\}$, where $\{\eta_{\cdot j}, \kappa_{\cdot j}\}$ are recycled from $h^{(1)}_{nuisance}$ but $\{\eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ are new nuisance functions after reparameterization (Table 1). Here, s is the probability of the cluster-specific assignment conditional on the all baseline information as well as post-baseline mediator information from that cluster; $\kappa^{\star}_{\cdot j}$ is the probability of $M_{\cdot j}$ conditional on assignment, baseline information, as well as post-baseline mediator information from the remaining same-cluster members; $\eta^{\star}_{\cdot j}$ and $\eta^{\dagger}_{\cdot j}$ are two conditional expectations for the (weighted) responses $\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)$ and $\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\frac{\kappa_{\cdot j}(a, M_{\cdot j}, \boldsymbol{C}, N)}{\kappa^{\star}_{\cdot j}(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}$, respectively.

This choice of parameterization in Proposition 1 can play an important role in operationalizing the semiparametric estimators based on the EIFs. First, the EIFs under reparameterization now only include one-dimensional integrals (rather than multi-dimensional ones), which can be conveniently obtained from existing numerical integration routines (such as the quadrature-based method or Monte Carlo integration). Second, all components in $h^{(2)}_{nuisance}$ are one-dimensional conditional density or expectations; this enables the specification of familiar statistical models, and dispenses with the specification of an association model required for a multivariate density function. Because nuisance functions in $h^{(2)}_{nuisance}$ also include variables with dimensions that vary across clusters, cluster-specific summary functions are recommended to complete model specification. For example, $s(a, \boldsymbol{m}, \boldsymbol{c}, n)$ can be estimated by regressing A_i on $\boldsymbol{V}_i$, N_i, and cluster-level summary functions of $\boldsymbol{X}_i$ and $\boldsymbol{M}_i$ (for example, $\overline{\boldsymbol{X}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \boldsymbol{X}_{ij}$ and $\overline{\boldsymbol{M}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \boldsymbol{M}_{ij}$). Estimation of $\kappa^{\star}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)$ is similar to $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$, except for the adjustment of a summary function of $\boldsymbol{M}_{i(-j)}$ (for example, $\frac{1}{N_i - 1}\sum_{l=1, l\neq j}^{N_i} M_{il}$). Finally, $\widehat{\eta}^{\star}_{\cdot j}(a, a^*, \boldsymbol{c}, n)$ can be obtained by regressing $\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ on A, $\boldsymbol{V}_i$, N_i and a summary function of $\boldsymbol{X}_i$, and $\widehat{\eta}^{\dagger}_{\cdot j}(a, a^*, m_{\cdot j}, \boldsymbol{c}, n)$ can be similarly obtained by regressing $\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)\frac{\widehat{\kappa}_{ij}(a, M_{ij}, \boldsymbol{C}_i, N_i)}{\widehat{\kappa}^{\star}_{ij}(a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)}$ on A_i, M_{ij}, $\boldsymbol{V}_i$, N_i, and a summary function of $\boldsymbol{X}_i$.

We notice that the nuisance functions in $h^{(2)}_{nuisance}$ are not always variationally independent. For example, $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}\}$ are variationally dependent since $\eta_{\cdot j}$ is nested in $\eta^{\star}_{\cdot j}$ and $\eta^{\dagger}_{\cdot j}$; $\{\kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ are also variationally dependent since both $\kappa_{\cdot j}$ and $\kappa^{\star}_{\cdot j}$ are conditional probabilities of $M_{\cdot j}$ and s is partially determined by $\{\kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$. Therefore, when parametric models are used for $h^{(2)}_{nuisance}$, it is important to ensure compatibility among the specification of parametric models for each nuisance function. When the mediator is continuous, we offer some practical strategies in the Supplementary Material for specifying parametric models of $h^{(2)}_{nuisance}$ that ensures compatibility under mild approximations.

Proposition 1 motivates alternative semiparametric estimators for assessing causal mediation in cluster-randomized experiments; we denote the new estimators by $\widehat{\zeta}^{\text{eif}_2\text{-par}}$, for $\zeta \in \{\theta_C(a, a^*), \theta_I(a, a^*), \tau_C, \tau_I\}$. The estimators have the same formulation as (7), but now $\widehat{\psi}_\theta(a, a^*, \boldsymbol{O})$ and $\widehat{\psi}_\tau(\boldsymbol{O})$ are obtained based on $\widehat{h}^{(2)}_{nuisance}$. Theorem 3 summarize the asymptotic properties of the two sets of semiparametric estimators, provided that all nuisance functions in $\widehat{h}^{(1)}_{nuisance}$ and $\widehat{h}^{(2)}_{nuisance}$ are obtained by parametric working models with finite-dimensional parameters.

Theorem 3. *(Double robustness and local efficiency) Suppose that the nuisance functions are estimated via parametric working models, Assumptions 1–4 hold for estimation of $\theta_V(a, a^*)$ ($V \in \{I, C\}$), Assumptions 1–5 hold for estimation of τ_V, and regularity conditions in Supplementary Material hold. The following results hold for both $V \in \{I, C\}$, and the expressions of each asymptotic variance is presented in the Supplementary Material.*

(i) If $\{\kappa_{\cdot j}, \mathcal{C}\}$ or $\eta_{\cdot j}$ is correctly specified, $\sqrt{K}\left\{\widehat{\theta}^{eif_1\text{-}par}_V(a, a^*) - \theta_V(a, a^*)\right\} \xrightarrow{d} N(0, \Sigma^{eif_1\text{-}par}_{\theta_V(a,a^*)})$; if $\{\eta_{\cdot j}, \kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified, $\Sigma^{eif_1\text{-}par}_{\theta_V(a,a^*)} = \mathbb{E}[\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})^2]$ achieves the efficiency lower bound for estimating $\theta_V(a, a^*)$. If $\{\kappa_{\cdot j}, \mathcal{C}\}$ or $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ are correctly specified, then $\sqrt{K}(\widehat{\tau}^{eif_1\text{-}par}_V - \tau_V) \xrightarrow{d} N(0, \Sigma^{eif_1\text{-}par}_{\tau_V})$; if $\{\eta_{\cdot j}, \kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified, $\Sigma^{eif_1\text{-}par}_{\tau_V} = \mathbb{E}[\mathcal{D}_{\tau_V}(\boldsymbol{O})^2]$ achieves the efficiency lower bound for estimating τ_V.

(ii) If $\eta_{\cdot j}$ or s is correctly specified, then $\sqrt{K}\left\{\widehat{\theta}^{eif_2\text{-}par}_V(a, a^*) - \theta_V(a, a^*)\right\} \xrightarrow{d} N(0, \Sigma^{eif_2\text{-}par}_{\theta_V(a,a^*)})$; if $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, s\}$ are correctly specified, $\Sigma^{eif_2\text{-}par}_{\theta_V(a,a^*)} = \mathbb{E}[\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})^2]$ achieves the efficiency

lower bound for estimating $\theta_V(a, a^*)$*. If* $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ *or* $\{\kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ *are correctly specified, then* $\sqrt{K}(\widehat{\tau}_V^{\textit{eif}_2\textit{-par}} - \tau_V) \xrightarrow{d} N(0, \Sigma_{\tau_V}^{\textit{eif}_2\textit{-par}})$*; if* $\{\eta_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ *are correctly specified,* $\Sigma_{\tau_V}^{\textit{eif}_2\textit{-par}} = \mathbb{E}[\mathcal{D}_{\tau_V}(\boldsymbol{O})^2]$ *achieves the efficiency lower bound for estimating* τ_V*.*

Theorem 3 indicates that the two sets of semiparametric estimators of $\theta_V(a, a^*)$ (i.e., $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a, a^*)$ and $\widehat{\theta}_V^{\text{eif}_2\text{-par}}(a, a^*)$) are doubly robust in the sense that they are $\sqrt{K}$-consistent and asymptotically normal (CAN) when one of two separate sets of parametric working models are correctly specified, but not necessarily both. In contrast to $\theta_V(a, a^*)$, the two semiparametric estimators of τ_V are *conditional* doubly robust requiring that $\kappa_{\cdot j}$ is correctly specified. Specifically, when $\kappa_{\cdot j}$ is correct, $\widehat{\tau}_V^{\text{eif}_1\text{-par}}$ is CAN if $\mathcal{C}$ or $\eta_{\cdot j}$ is correctly specified whereas $\widehat{\tau}_V^{\text{eif}_2\text{-par}}$ is CAN if $\eta_{\cdot j}$ or $\{\kappa^\star_{\cdot j}, s\}$ are correctly specified. Interestingly, we observe that the consistency of $\{\widehat{\theta}_V^{\text{eif}_2\text{-par}}(a, a^*), \widehat{\tau}_V^{\text{eif}_2\text{-par}}\}$ does not depend on the specification of $\{\eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}\}$, but their local efficiency necessitates the correct specification of $\{\eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}\}$. Finally, since each asymptotic variance expression in Theorem 3 has a rather complicated form, we recommend nonparametric cluster bootstrap (Field and Welsh, 2007) for inference. Specifically, we resample the K clusters with replacement (but without further within-cluster resampling) and re-compute each estimator for a large number of bootstrap replications. Then, the $\alpha/2$ and $(1-\alpha/2)$-quantiles of the bootstrap distribution are used to form a $(1-\alpha) \times 100\%$ confidence interval for each causal mediation estimands.

Remark 2. *(Bonus robustness) In addition to the robustness properties in Theorem 3, the proposed estimators remain consistent under an additional condition when* $\eta_{\cdot j}$ *follows a linear regression without mediator-by-mediator and mediator-by-covariate interactions. This corresponds to a data generating distribution compatible with* $\eta_{\cdot j}(A, \boldsymbol{M}, \boldsymbol{C}, N) = \beta_0(A, \boldsymbol{C}, N) + \sum_{l=1}^{N} \beta_{1,l}(A, N) M_{\cdot j}$*, where* $\beta_0(A, \boldsymbol{C}, N)$ *and* $\beta_{1,l}(A, N)$ *are unknown functions of* $\{A, \boldsymbol{C}, N\}$ *and* $\{A, N\}$*. Suppose the mean of* $\kappa_{\cdot j}$ *is modeled by a generalized linear model (GLM) with* $\upsilon(\mathbb{E}[M_{\cdot j}|A, N]) = \boldsymbol{U}^T\boldsymbol{\alpha}$*, where* υ *is a link function and* $\boldsymbol{U}$ *is the design matrix of* $\{A, N\}$*. If (i) the GLM for* $\mathbb{E}[M_{\cdot j}|A, N]$ *is correctly specified, and (ii)* $\eta_{\cdot j}$

is correctly specified, then $\{\widehat{\theta}_V^{d}(a,a^*), \widehat{\tau}_V^{d}\}$ *for* $d \in \{$*mf-par*, *eif*$_1$*-par*$\}$*, are consistent, irrespective of whether the remainder nuisance functions are correctly specified. In addition, for* $\{\widehat{\theta}_V^{eif_2\text{-}par}(a,a^*), \widehat{\tau}_V^{eif_2\text{-}par}\}$*, if we further fit* $\eta^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$ *with a linear regression with no interaction terms between* $M_{\cdot j}$ *and* $\boldsymbol{C}$*, then they are also consistent under (i) and (ii) regardless of whether the linear regression of* $\eta^{\dagger}_{\cdot j}$ *is correct or not. Because* $\mathbb{E}[M_{\cdot j}|A,N]$ *only includes two predictors, risk of misspecification of* $\mathbb{E}[M_{\cdot j}|A,N]$ *may be minimized by including flexible spline terms and interaction terms in the GLM. In this case, all semiparametric estimators for* $\theta_V(a,a^*)$ *and* τ_V *are consistent as long as* $\eta_{\cdot j}$ *is correctly specified by a linear regression, regardless of whether other associated nuisance functions are correctly specified.*

4.3 Leveraging machine learning to obtain efficient estimators

To potentially improve the semiparametric estimators under parametric model misspecification, we extend $\widehat{\zeta}^{\text{eif}_1\text{-par}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-par}}$ (for $\zeta \in \{\theta_C(a,a^*), \theta_I(a,a^*), \tau_C, \tau_I\}$) by leveraging data-adaptive machine learners to estimate the nuisance functions. This leads to two machine learning estimators $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$. As $h^{(2)}_{nusiance}$ only involves conditional expectations or conditional densities of univariate variables, several off-the-shelf machine learners can be employed to obtain their estimates (see Phillips et al. (2023) for a list concrete machine learners and a guide on optimizing their performance through the Super Learner). Using machine learners to estimate $h^{(1)}_{nusiance}$ may be more challenging, because $h^{(1)}_{nusiance}$ involves two multi-dimensional conditional densities $\{\kappa, \kappa_{\cdot(-j)}\}$. In this case, similar to Section 4.1, we parameterize $h^{(1)}_{nusiance}$ into $\{\eta_{\cdot j}, \kappa^c, \kappa_{\cdot j}, \kappa^c_{\cdot(-j)}\}$ based on a multivariate copula $\mathcal{C}$ given in (4). Then, $\widehat{\eta}_{\cdot j}$ and $\widehat{\kappa}_{\cdot j}$ can still be based on machine learners designed for conditional expectation and conditional density. For modeling $\mathcal{C}$, we consider the parametric Gaussian copula (5), where $\widehat{\boldsymbol{\rho}}$ is obtained by the pseudo-likelihood approach except that the machine learning estimator $\widehat{\kappa}_{\cdot j}$ is plugged in the pseudo-likelihood; the full details are found in the Supplementary Material.

When data-adaptive machine learners are applied, cross-fitting is necessary to control the empirical process term to be asymptotically negligible (Chernozhukov et al., 2018). We outline the steps to calculate $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ under cross-fitting, and similar procedures apply to $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$. Under cross-fitting, we partition the index set $\{1,\ldots,K\}$ into R non-overlapping parts $\{\mathcal{S}_1,\cdots,\mathcal{S}_R\}$ of approximately equal size. For each $r\in\{1,\ldots,R\}$, denote $\mathcal{S}_r$ as the validation sample and $\mathcal{S}_{-r}=\{1,\ldots,K\}\setminus\mathcal{S}_r$ be the training sample. For all i in $\mathcal{S}_r$, we calculate $\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i)$ and $\widehat{\psi}_\tau(\boldsymbol{O}_i)$ with nuisance functions learned from the training sample $\mathcal{S}_{-r}$. We then repeat the previous step from $\mathcal{S}_1$ to $\mathcal{S}_R$ to obtain all $\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i)$ and $\widehat{\psi}_\tau(\boldsymbol{O}_i)$ for $i=1\ldots,K$. Finally, we substitute all $\widehat{\psi}_\theta(a,a^*;\boldsymbol{O}_i)$ and $\widehat{\psi}_\tau(\boldsymbol{O}_i)$ into (7) to obtain $\widehat{\theta}_C^{\text{eif}_1\text{-ml}}(a,a^*)$, $\widehat{\theta}_I^{\text{eif}_1\text{-ml}}(a,a^*)$, $\widehat{\tau}_C^{\text{eif}_1\text{-ml}}$, and $\widehat{\tau}_I^{\text{eif}_1\text{-ml}}$. The asymptotic properties of $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$ are provided in Theorem 4. As an additional notation, we let $\lambda_K(\cdot)$ denote a mapping from a nuisance function estimate to its convergence rate under the $L_2(P)$-norm, where $P=P_{\boldsymbol{O}}$ represents the true distribution of observed data.

Theorem 4. *(Efficiency with machine learning estimators) Suppose that Assumptions 1–4 hold for estimation of $\theta_V(a,a^*)$ ($V\in\{I,C\}$), Assumptions 1–5 hold for estimation of τ_V, and the nuisance functions are estimated based on machine learners with cross-fitting. The following results hold for both $V\in\{I,C\}$.*

(i) $\widehat{\theta}_V^{\mathit{eif}_1\text{-}\mathit{ml}}(a,a^*)$ is consistent if either $\widehat{\kappa}^c$ or $\widehat{\eta}_{\cdot j}$ is consistent in $L_2(P)$-norm; furthermore, if both $\{\widehat{\kappa}^c,\widehat{\eta}_{\cdot j}\}$ are consistent with $\lambda_K(\widehat{\kappa}^c)\lambda_K(\widehat{\eta}_{\cdot j})=o(K^{-1/2})$, then $\sqrt{K}(\widehat{\theta}_V^{\mathit{eif}_1\text{-}\mathit{ml}}(a,a^*)$ $-\ \theta_V(a,a^*))\xrightarrow{d}N(0,\mathbb{E}[\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})^2])$. $\widehat{\tau}_V^{\mathit{eif}_1\text{-}\mathit{ml}}$ is consistent if either $\{\widehat{\kappa}_{\cdot j},\widehat{\eta}_{\cdot j}\}$ or $\{\widehat{\kappa}_{\cdot j},\widehat{\kappa}^c_{\cdot(-j)},\widehat{\kappa}^c\}$ are consistent in $L_2(P)$-norm; furthermore, if $\{\widehat{\kappa}_{\cdot j},\widehat{\eta}_{\cdot j},\widehat{\kappa}^c_{\cdot(-j)},\widehat{\kappa}^c\}$ are consistent with $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j})+\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^c_{\cdot(-j)})+\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^c)+\lambda_K(\widehat{\kappa}^c_{\cdot(-j)})\lambda_K(\widehat{\kappa}_{\cdot j})=o(K^{-1/2})$, then $\sqrt{K}\left(\widehat{\tau}_V^{\mathit{eif}_1\text{-}\mathit{ml}}-\tau_V\right)\xrightarrow{d}N(0,\mathbb{E}[\mathcal{D}_{\tau_V}(\boldsymbol{O})^2])$.

(ii) $\widehat{\theta}_V^{\mathit{eif}_2\text{-}\mathit{ml}}(a,a^*)$ is consistent if either $\widehat{\eta}_{\cdot j}$ or $\widehat{s}$ is consistent in $L_2(P)$-norm; furthermore, if $\{\widehat{\eta}_{\cdot j},\widehat{\eta}^\star_{\cdot j},\widehat{s}\}$ are consistent with $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{s})=o(K^{-1/2})$, then $\sqrt{K}(\widehat{\theta}_V^{\mathit{eif}_2\text{-}\mathit{ml}}(a,a^*)-$ $\theta_V(a,a^*))\xrightarrow{d}N(0,\mathbb{E}[\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})^2])$. $\widehat{\tau}_V^{\mathit{eif}_2\text{-}\mathit{ml}}$ is consistent if either $\{\widehat{\kappa}_{\cdot j},\widehat{\eta}_{\cdot j}\}$ or $\{\widehat{\kappa}_{\cdot j},\widehat{\kappa}^\star_{\cdot j},\widehat{s}\}$

are consistent in $L_2(P)$*-norm; furthermore, if* $\{\widehat{\eta}_{\cdot j}, \widehat{\eta}^{\dagger}_{\cdot j}, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^{\star}_{\cdot j}, \widehat{s}\}$ *are consistent with* $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^{\star}_{\cdot j}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{s}) + \lambda_K(\widehat{\eta}^{\dagger}_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j}) = o(K^{-1/2})$*, then* $\sqrt{K}(\widehat{\tau}_V^{eif_1\text{-}ml} - \tau_V) \xrightarrow{d} N(0, \mathbb{E}[\mathcal{D}_{\tau_V}(\boldsymbol{O})^2])$.

The robustness properties of $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$ echo their doubly robust counterparts when all nuisances are specified by parametric models. Furthermore, Theorem 4 also reveals that $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$ are also asymptotically normal and semiparametrically efficient when the nuisance functions involved in each estimator are consistent in $L_2(P)$-norm and satisfy mild conditions for convergence rate. In practice, a $o_p(K^{-1/4})$-type convergence rate among nuisance functions is sufficient to ensure all $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$ to be consistent, asymptotically normal, and efficient. For inference, we use the empirical variance of EIF under cross-fitting to construct the uncertainty estimators; for example, $\text{Var}(\widehat{\tau}_C^{\text{eif}_1\text{-ml}})$ and $\text{Var}(\widehat{\tau}_I^{\text{eif}_1\text{-ml}})$ can be estimated by $\frac{1}{K}\mathbb{P}_K\left[\widehat{\mathcal{D}}_{\tau_C}(\boldsymbol{O})^2\right] = \frac{1}{K^2}\sum_{i=1}^K \left\{\widehat{\psi}_\tau(\boldsymbol{O}_i) - \widehat{\tau}_C^{\text{eif}_1\text{-ml}}\right\}^2$ and $\frac{1}{K}\mathbb{P}_K\left[\widehat{\mathcal{D}}_{\tau_I}(\boldsymbol{O})^2\right] = \frac{1}{K^2}\sum_{i=1}^K \left\{\frac{N_i}{\overline{N}}[\widehat{\psi}_\tau(\boldsymbol{O}_i) - \widehat{\tau}_I^{\text{eif}_1\text{-ml}}]\right\}^2$, respectively.

4.4 Stabilization to improve finite-sample performance

All semiparametric and machine learning estimators involve density ratios, $w^{(1)}$ and $w^{(2)}$, regardless of whether the reparameterization in Proposition 1 is considered. When positivity of $f_{\boldsymbol{M}|A,\boldsymbol{C},N}(\boldsymbol{m}|a,\boldsymbol{c},n) > 0$ in Assumption 4 is empirically violated or the nuisance functions in the density ratios are inconsistently estimated, then $\widehat{w}^{(1)}$ and $\widehat{w}^{(2)}$ may be highly variable and a stabilization procedure may improve the finite-sample performance of our estimators. Following a similar spirit of Tchetgen and Shpitser (2012), Miles et al. (2020), and Zhou (2022), we stabilize the estimators by customizing each estimating equations so that the terms involving $w^{(1)}$ and $w^{(2)}$ are empirically set equal to zero. We summarize the steps to obtain stabilized $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a, a^*)$ $(V \in \{C, I\})$ below.

1. Obtain $\widehat{h}^{(1)}_{nuisance} = \{\widehat{\eta}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^c_{\cdot(-j)}\}$ based on the original procedure.
2. Define $\varrho(N) = \frac{1}{N}$ and 1 if $V = C$ and I respectively. For all clusters with treatment

$A_i = a$, fit a GLM for the conditional mean of Y_{ij}, using $\upsilon(\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i))$ as an offset term with $\varrho(N_i)\widehat{w}^{(1)}(a, a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the only covariate (without an intercept), where $\upsilon(\cdot)$ is a canonical link function. Then, update $\widehat{\eta}^{\text{stab}}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \upsilon^{-1}(\upsilon(\widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)) + \widehat{\beta}\varrho(n)\widehat{w}^{(1)}(a, a^*, \boldsymbol{m}, \boldsymbol{c}, n))$, where $\widehat{\beta}$ is the estimated regression coefficient in the GLM.

3. For $i = 1, \ldots, K$, calculate $\widehat{\psi}^{\text{stab}}_{\theta}(a, a^*; \boldsymbol{O}_i)$ by

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\left\{\frac{\mathbb{I}(A = a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left[\widehat{\eta}^{\text{stab}}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i) - \widehat{u}^{(1)}_{ij}(a, a^*, \boldsymbol{C}_i, N_i)\right] + \widehat{u}^{(1)}_{ij}(a, a^*, \boldsymbol{C}_i, N_i)\right\}$$

based on the updated estimates $\widehat{h}^{(1)}_{nuisance} = \{\widehat{\eta}^{\text{stab}}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^c_{\cdot(-j)}\}$. Then, substitute $\widehat{\psi}^{\text{stab}}_{\theta}(a, a^*; \boldsymbol{O}_i)$ into (7) to obtain the stabilized $\widehat{\theta}^{\text{eif}_1\text{-par}}_V(a, a^*)$.

Following a similar strategy, we can construct stabilized estimators for $\widehat{\tau}^{\text{eif}_1\text{-par}}_C$ and $\widehat{\tau}^{\text{eif}_1\text{-par}}_I$, as well as $\widehat{\zeta}^{\text{eif}_2\text{-par}}$ with $\zeta \in \{\theta_C(a, a^*), \theta_I(a, a^*), \tau_C, \tau_I\}$. In addition, the stabilized procedures can be generalized to the machine learning estimators under cross-fitting, where details of all stabilized estimators are summarized in Supplementary Material.

4.5 Estimating the causal mediation effect estimands

The mediation effects can be immediately obtained by directly plugging the estimates for $\theta_V(a, a^*)$ and τ_V ($V \in \{C, I\}$) into the estimands definitions (such as those in (2) and (3)). Specifically, for any approach $d \in \{\text{mf-par}, \text{eif}_1\text{-par}, \text{eif}_2\text{-par}, \text{eif}_2\text{-ml}, \text{eif}_2\text{-ml}\}$, plug-in estimates of the NIE_V, NDE_V, SME_V, and IME_V can be calculated by $\widehat{\text{NIE}}^d_V = g\left(\widehat{\theta}^d_V(1,1), \widehat{\theta}^d_V(1,0)\right)$, $\widehat{\text{NDE}}^d_V = \left(\widehat{\theta}^d_V(1,0), \widehat{\theta}^d_V(0,0)\right)$, $\widehat{\text{SME}}^d_V = g\left(\widehat{\theta}^d_V(1,1), \widehat{\tau}^d_V\right)$, and $\widehat{\text{IME}}^d_V = g\left(\widehat{\tau}^d_V, \widehat{\theta}^d_V(1,0)\right)$. In the Supplementary Material, we also summarize the asymptotic properties of each NIE, NDE, SME and IME estimator, when the nuisance functions are estimated through either parametric regression models or data-adaptive machine learners. The properties match those proved in Theorems 3 and 4.

5 Simulation studies

We conduct simulations to illustrate the performance characteristics of the proposed methods. We simulate 1000 cluster-randomized experiments, each with $K = 100$ clusters, where the treatment is randomized based on a Bernoulli distribution $A_i \sim \text{Bernoulli}(0.5)$. The data generation process for the observed data $\boldsymbol{O}_i = \{N_i, \boldsymbol{C}_i, A_i, \boldsymbol{M}_i, \boldsymbol{Y}_i\}$ are detailed in Supplementary Material, where the mediator and outcome are both considered as continuous variables to align with our settings in application studies. We consider the cluster- and individual-average NIE and SME estimands on the mean difference scale. The true mediation effect estimands are $\text{NIE}_C = 0.75$, $\text{NIE}_I = 0.82$, $\text{SME}_C = 0.25$, and $\text{SME}_I = 0.27$. For each $\Delta \in \{\text{NIE}_C, \text{NIE}_I, \text{SME}_C, \text{SME}_I\}$, we compare the performance among nine estimators proposed in Section 4: $\widehat{\Delta}^{\text{mf-par}}, \widehat{\Delta}^{\text{eif}_1\text{-par-o}}, \widehat{\Delta}^{\text{eif}_1\text{-par-s}}, \widehat{\Delta}^{\text{eif}_2\text{-par-o}}, \widehat{\Delta}^{\text{eif}_2\text{-par-s}}, \widehat{\Delta}^{\text{eif}_1\text{-ml-o}}$, $\widehat{\Delta}^{\text{eif}_1\text{-ml-s}}, \widehat{\Delta}^{\text{eif}_2\text{-ml-o}}$, and $\widehat{\Delta}^{\text{eif}_2\text{-ml-s}}$. For all EIF-based estimators, we add the superscript '-o' and '-s' to further distinguish the original EIF-based estimator without stabilization and the EIF-based estimators after stabilization in Section 4.4; for example, $\widehat{\Delta}^{\text{eif}_1\text{-par-o}}$ is the original $\widehat{\Delta}^{\text{eif}_1\text{-par}}$ and $\widehat{\Delta}^{\text{eif}_1\text{-par-s}}$ is its counterpart after stabilization. We also evaluate the 95% confidence interval coverage associated with each estimator, based on cluster bootstrap for parametric estimators and the proposed variance estimator for the machine-learning estimators. To improve the finite-sample coverage of the machine learning estimators, we follow the suggestion from Wang et al. (2023) and apply a degree of freedom adjustment by multiplying the variance estimator by $\frac{K}{K-3}$, where 3 is the number of adjusted baseline covariates $\{V_i, \boldsymbol{X}_i, N_i\}$; moreover, we improve the finite-sample coverage by using a t-distribution with $K - 3$ degrees of freedom instead of a normal distribution.

Among the five semiparametric estimators, $\{\widehat{\Delta}^{\text{mf-par}}, \widehat{\Delta}^{\text{eif}_1\text{-par-o}}, \widehat{\Delta}^{\text{eif}_1\text{-par-s}}\}$ are constructed based on parametric models for $\{\eta_{\cdot j}, \kappa_{\cdot j}, \mathcal{C}\}$, and $\{\widehat{\Delta}^{\text{eif}_2\text{-par-o}}, \widehat{\Delta}^{\text{eif}_2\text{-par-s}}\}$ are constructed based on parametric models for $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$. Specifically, we consider a Gaussian copula with an exchangeable association structure for $\mathcal{C}$, a logistic regression for s, a linear regression for $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}\}$, and a linear regression with a constant Gaussian error for $\{\kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ (see Supplementary Material for details). To demonstrate the robustness of the proposed estimators, we also evaluate each estimator when a subset of parametric working models are misspecified; in these cases the working regression model is obtained by removing all baseline covariates and the misspecified copula is obtained by fixing $\rho = 0$ (the mediators in the same cluster are incorrectly assumed to be independent). For $\{\widehat{\Delta}^{\text{mf-par}}, \widehat{\Delta}^{\text{eif}_1\text{-par-o}}, \widehat{\Delta}^{\text{eif}_1\text{-par-s}}\}$ that require specification of $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider 5 scenarios including: (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ are correctly specified, (b) only $\{\kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified, (c) only $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ are correctly specified, (d) only $\{\eta_{\cdot j}, \mathcal{C}\}$ are correctly specified, and (e) only $\mathcal{C}$ is correctly specified. For $\{\widehat{\Delta}^{\text{eif}_2\text{-par-o}}, \widehat{\Delta}^{\text{eif}_2\text{-par-s}}\}$ that requires specification of $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we also consider 5 scenarios including: (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ are correctly specified, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ are correctly specified, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ are correctly specified, (4) only $\{s, \eta_{\cdot j}\}$ are correctly specified, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ are misspecified.

We consider a five-fold cross-fitting for the four machine learning estimators. Recall that $\{\widehat{\Delta}^{\text{eif}_1\text{-ml-o}}, \widehat{\Delta}^{\text{eif}_1\text{-ml-s}}\}$ and $\{\widehat{\Delta}^{\text{eif}_2\text{-ml-o}}, \widehat{\Delta}^{\text{eif}_2\text{-ml-s}}\}$ require estimation of $\{\eta_{\cdot j}, \kappa_{\cdot j}, \kappa, \kappa_{\cdot(-j)}\}$ and $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, respectively; among them, $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}\}$ are estimated based on the Super Learner (Phillips et al., 2023) with random forest and generalized linear model libraries. To estimate $\{\kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$, we consider two additive models $M_{\cdot j} = \mathbb{E}[M_{\cdot j}|A, \boldsymbol{C}, N] + \epsilon_{M_{\cdot j}}$ and $M_{\cdot j} = \mathbb{E}[M_{\cdot j}|A, \boldsymbol{M}_{\cdot(-j)}, \boldsymbol{C}, N] + \epsilon^{\star}_{M_{\cdot j}}$, where the conditional mean structure is estimated through Super Learner with random forest and linear regression libraries, and the error terms $\{\epsilon_{M_{\cdot j}}, \epsilon^{\star}_{M_{\cdot j}}\}$ are assumed normally distributed with variance estimated via the empirical variance of the residuals. To estimate $\{\kappa, \kappa_{\cdot(-j)}\}$, we consider their copula representation $\{\kappa^c, \kappa^c_{\cdot(-j)}\}$ (a Gaussian copula with an exchangeable association structure). Similar to the estimators based on parametric working models, we investigate the performance of the machine learning estimators under misspecification under Scenarios (a)–(e) for $\{\widehat{\Delta}^{\text{eif}_1\text{-ml-o}}, \widehat{\Delta}^{\text{eif}_1\text{-ml-s}}\}$ and Scenarios (1)–(5) for $\{\widehat{\Delta}^{\text{eif}_2\text{-ml-o}}, \widehat{\Delta}^{\text{eif}_2\text{-ml-s}}\}$. For correctly specified nuisance function in each scenario, we supply the Super Learner with a feature matrix that is used in the correctly specified parametric working model; otherwise, all baseline covariates are removed from the feature matrix in each misspecification scenario.

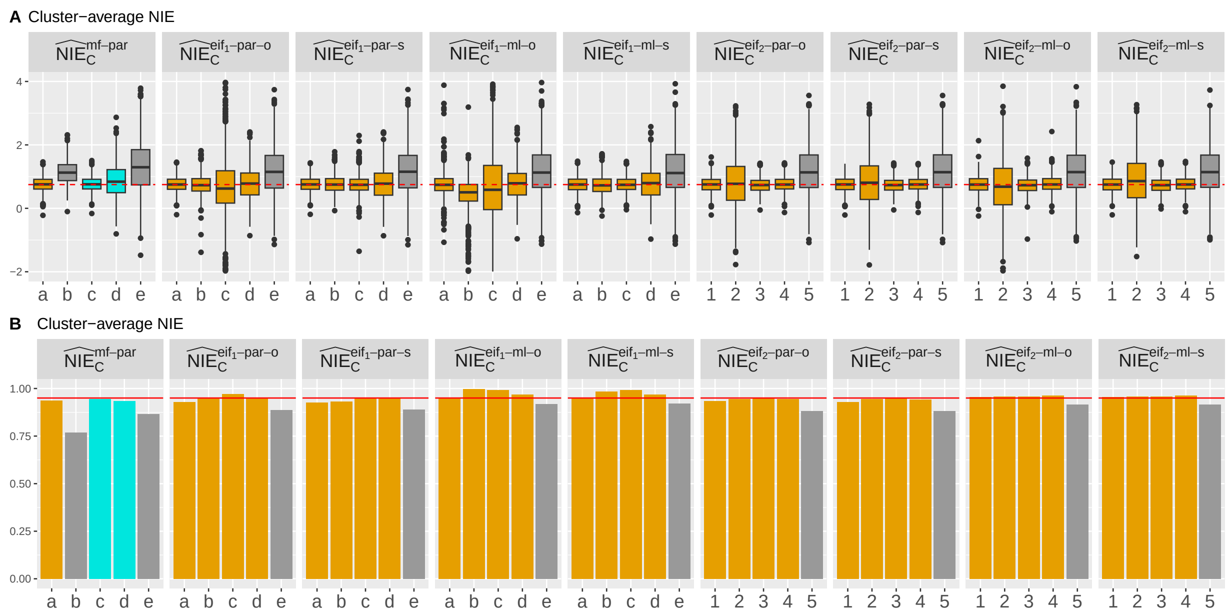

Figure 2: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of NIE_C. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ incorrect.

Figure 2(A) presents boxplots for estimators of NIE_C. According to Section 4, $\widehat{\text{NIE}}_C^{\text{mf-par}}$ is consistent in Scenario (a), $\{\widehat{\text{NIE}}_C^{\text{eif}_1\text{-par-o}}, \widehat{\text{NIE}}_C^{\text{eif}_1\text{-par-s}}, \widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-o}}, \widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-s}}\}$ are consistent in Scenarios (a)–(d), and $\{\widehat{\text{NIE}}_C^{\text{eif}_2\text{-par-o}}, \widehat{\text{NIE}}_C^{\text{eif}_2\text{-par-s}}, \widehat{\text{NIE}}_C^{\text{eif}_2\text{-ml-o}}, \widehat{\text{NIE}}_C^{\text{eif}_2\text{-ml-s}}\}$ are consistent in Scenarios (1)–(4). In addition, $\widehat{\text{NIE}}_C^{\text{mf-par}}$ is also consistent in Scenarios (c) and (d) when $\kappa_{\cdot j}$ and/or $\mathcal{C}$ are misspecified, because the true outcome is generated from a linear model without mediator-by-mediator and mediator-by-covariate interactions and $\widehat{\mathbb{E}}[M_{\cdot j}|A, N]$ induced from $\widehat{\kappa}_{\cdot j}$ is the correct model for $\mathbb{E}[M_{\cdot j}|A, N]$ (Remark 2). We observe that all estimators behave as expected; that is, they concentrate around the true estimand when the requisite nuisance functions are correctly specified, but may deviate from the truth otherwise. Furthermore, stabilized estimators often improve the efficiency over their unstabilized counterparts. For example, $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-par-o}}$ becomes more variable than $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-par-s}}$ when the copula $\mathcal{C}$ is misspecified (Scenario (c)). In Scenario (b) when $\eta_{\cdot j}$ is misspecified, $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-s}}$ presents small variance with negligible bias but $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-o}}$ exhibits more noticeable finite-sample bias (additional unreported simulations reveal that $K \geq 500$ clusters is sufficient to remove the finite-sample bias of $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-o}}$). Figure 2(B) presents the 95% confidence interval coverage among each estimator of NIE_C. In general, all methods provide close-to-nominal coverage in scenarios that they are expected to be consistent.

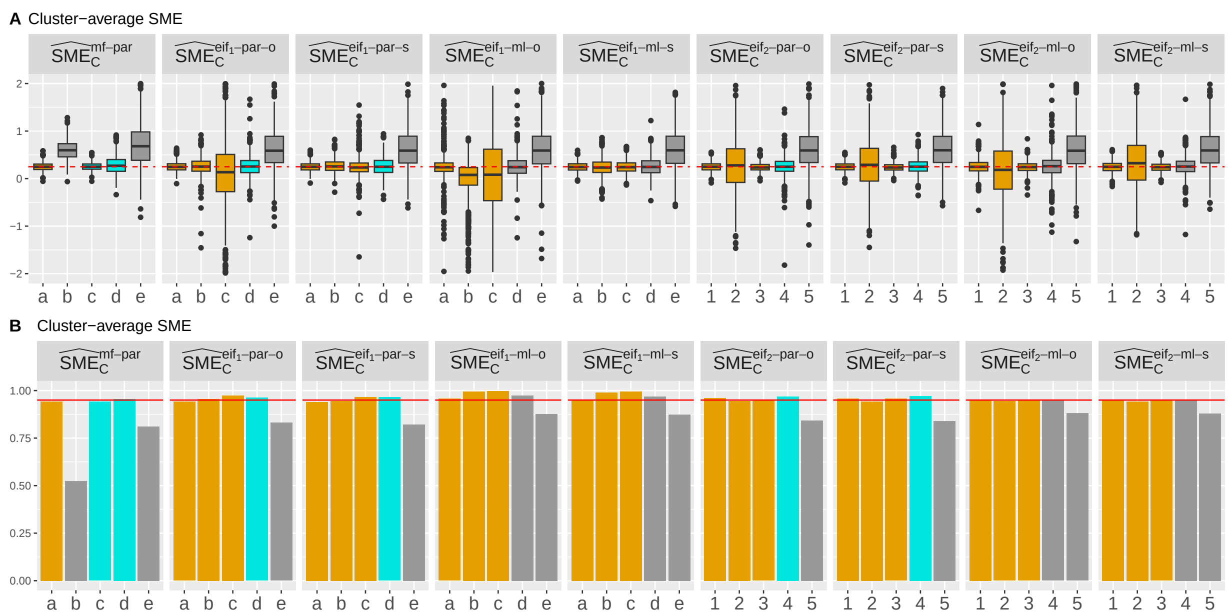

Figure 3: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of SME_C. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ incorrect.

Figure 3(A) presents boxplots for estimators of SME_C. Consistent with our theory, $\widehat{\text{SME}}_C^{\text{mf-par}}$ leads to minimal bias in Scenario (a), $\{\widehat{\text{SME}}_C^{\text{eif}_1\text{-par-o}}, \widehat{\text{SME}}_C^{\text{eif}_1\text{-par-s}}\}$ lead to minimal bias in Scenarios (a)–(c), and $\{\widehat{\text{SME}}_C^{\text{eif}_1\text{-par-o}}, \widehat{\text{SME}}_C^{\text{eif}_1\text{-par-s}}\}$ lead to minimal bias in Scenarios (1)–(3). Besides, in line with Remark 2, $\widehat{\text{SME}}_C^{\text{mf-par}}$ has minimal bias in Scenarios (c)–(d), $\{\widehat{\text{SME}}_C^{\text{eif}_1\text{-par-o}}, \widehat{\text{SME}}_C^{\text{eif}_1\text{-par-s}}\}$ have minimal bias in Scenarios (d) and $\{\widehat{\text{SME}}_C^{\text{eif}_2\text{-par-o}}, \widehat{\text{SME}}_C^{\text{eif}_2\text{-par-s}}\}$ have minimal bias in Scenarios (4). The bias and variability among the machine learning estimators are similar to their parametric counterparts, except that $\widehat{\text{SME}}_C^{\text{eif}_2\text{-ml-o}}$ exhibits finite-sample bias in Scenario (b). The stabilized estimators outperform their unstablized counterparts in terms of the bias and variability. The 95% confidence interval coverage among all estimators of SME_C is close-to-nominal when the respective point estimator is expected to be consistent (Figure 3(B)).

We have also investigated the performance for estimators of NIE_I and SME_I, and the results are qualitatively similar to estimators of NIE_C and SME_C. Finally, we have also evaluated the performance for each estimator in a smaller-sample setting with $K = 50$ clusters, and the patterns of results are similar. Details of these additional simulation results are reported in the Supplementary Material Figures 1–6.

6 Application to two cluster-randomized experiments

6.1 A cash-transfer program for child nutritional status

The *Red de Protección Social* pilot study (meaning "Social Safety Net" in Spanish) is a cluster-randomized experiment designed to investigate effectiveness of a cash-transfer program on a survey sample of households living in poverty across $K = 42$ *comarcas* (an administrative region) in Nicaragua (Charters et al., 2023). Randomization was performed on the *comarca* level with equal allocation; only households in treated *comarcas* were provided with conditional cash transfers. We aim to assess the role of household dietary diversity in mediating the treatment effect from the cash-transfer program on child nutritional status. Specifically, the mediator is measured by a household dietary diversity score (ranged between 0 and 12) and the outcome is measured by child height-by-age z-scores (mean -1.7, standard deviation 1.2), with a higher height-by-age z-score suggesting a better child nutritional status. In assessing causal mediation, we adjust for cluster size (mean 10.7 and standard deviation 4.8) and the following individual-level baseline covariates: mother's educational level, mother's literacy, and highest education in household. We restrict our analysis on children aged 6–35 months.

Table 2: Mediation of the effect of a cash-transfer program on child height-by-age z-score by household dietary divsersity score, the *Red de Protección Social* pilot in Nicaragua, 2000–2002. The numbers in parentheses are the associated 95% confidence interval.

Type	Estimator	NIE	NDE	SME	MP¶	SP¶
Cluster average	mf-par	0.17 (−0.17, 0.52)	0.17 (−0.26, 0.58)	0.11 (−0.23, 0.44)	50.9%	62.2%
	eif_1-par-o	0.16 (−0.21, 0.51)	0.16 (−0.26, 0.56)	−0.06 (−0.40, 0.44)	50.1%	–
	eif_1-par-s	0.19 (−0.24, 0.72)	0.12 (−0.49, 0.59)	0.00 (−0.39, 0.57)	60.3%	–
	eif_1-ml-o	0.14 (−0.22, 0.51)	0.20 (−0.26, 0.66)	−0.39 (−1.62, 0.84)	41.5%	–
	eif_1-ml-s	0.39 (0.14, 0.63)	0.05 (−0.35, 0.44)	0.22 (−0.01, 0.45)	89.4%	56.6%
	eif_2-par-o	0.19 (−0.23, 0.64)	0.13 (−0.34, 0.56)	0.02 (−0.53, 0.49)	58.3%	10.3%
	eif_2-par-s	0.24 (−0.32, 0.70)	0.08 (−0.47, 0.60)	0.05 (−0.43, 0.53)	75.3%	20.8%
	eif_2-ml-o	0.19 (−0.13, 0.51)	0.18 (−0.20, 0.56)	0.02 (−0.42, 0.46)	52.0%	10.7%
	eif_2-ml-s	0.20 (−0.10, 0.50)	0.24 (−0.16, 0.64)	−0.01 (−0.30, 0.28)	45.3%	–
Individual average	mf-par	0.16 (−0.16, 0.56)	0.14 (−0.25, 0.49)	0.10 (−0.21, 0.46)	53.4%	62.1%
	eif_1-par-o	0.18 (−0.17, 0.54)	0.13 (−0.25, 0.51)	0.01 (−0.32, 0.44)	58.9%	3.5%
	eif_1-par-s	0.26 (−0.22, 0.83)	0.05 (−0.46, 0.56)	0.14 (−0.28, 0.58)	83.9%	55.2%
	eif_1-ml-o	0.15 (−0.19, 0.49)	0.20 (−0.21, 0.61)	−0.27 (−1.33, 0.79)	42.9%	–
	eif_1-ml-s	0.26 (0.03, 0.48)	0.19 (−0.17, 0.55)	0.10 (−0.10, 0.30)	57.5%	37.9%
	eif_2-par-o	0.25 (−0.22, 0.69)	0.09 (−0.32, 0.57)	0.10 (−0.50, 0.68)	73.6%	41.5%
	eif_2-par-s	0.37 (−0.27, 0.77)	−0.03 (−0.40, 0.56)	0.23 (−0.40, 0.58)	–	62.3%
	eif_2-ml-o	0.27 (−0.06, 0.59)	0.12 (−0.30, 0.55)	0.15 (−0.31, 0.61)	68.2%	56.7%
	eif_2-ml-s	0.19 (−0.11, 0.50)	0.26 (−0.14, 0.66)	−0.02 (−0.28, 0.24)	42.7%	–

¶ 'MP' is the mediation proportion, defined as the ratio between the NIE and the total effect (i.e., NIE+NDE). 'SP' is the spillover proportion, defined as the ratio between the SME and NIE. We only present the MP and SP when their values lies between 0 and 1.

Table 2 presents estimates of the natural indirect effect, natural direct effect, and spillover mediation effect, defined based on both the cluster- and individual-level averages with a difference scale. Specifications of the parametric working models and machine learners follow those described in the simulation study and details are included Supplementary Material. Although all methods report a positive point estimate of the natural indirect effects, most of the associated 95% confidence intervals include zero, except that two stabilized estimators $\widehat{\text{NIE}}_C^{\text{eif}_1\text{-ml-s}} = 0.39$ (95% CI: [0.14,0.63]) and $\widehat{\text{NIE}}_I^{\text{eif}_1\text{-ml-s}} = 0.26$ (95% CI: [0.03,0.48]) suggest some evidence that the cash-transfer program significantly improved child height-by-age z-score through the household dietary diversity. This discrepancy across methods may be due to the cap in total sample size is relatively small (a total of 449 children) that limits the statistical efficiency of the analytic methods. In addition, the majority of estimates of the spillover mediation effect are close to 0 with 95% confidence interval including zero. Despite the overall similarity in the conclusions, we observe empirically that the stablization leads to efficiency gain as evidenced by a narrower confidence interval; for example, when machine learning models are used to estimate the nuisance functions, $\widehat{\text{SME}}_I^{\text{eif}_1\text{-ml-o}}$ is associated with a wide confidence interval $(-0.33, 0.79)$ but after stabilization a substantially narrower interval, $(-0.10, 0.30)$, is reported by $\widehat{\text{SME}}_I^{\text{eif}_1\text{-ml-s}}$.

6.2 A cognitive behavioral therapy for long-term opioid users

The Pain Program for Active Coping and Training study is a pragmatic, cluster-randomized experiment to examine a primary care-based cognitive behavior therapy intervention to improve chronic pain among long-term opioid users (DeBar et al., 2022). A total of $K = 106$ clusters of primary care providers were randomized with an equal allocation ratio to receive the cognitive behavior therapy intervention or usual care. On average, each cluster has $N = 8$ patients (standard deviation 2.3). In our analysis, the mediator is patient-reported pain impact at 6 months follow-up, assessed using a PEGS score (pain intensity and interference with enjoyment of life, general activity, and sleep). The outcome is defined as the patient-reported pain-related disability at 12 months follow-up, measured by a RMDQ score (Roland Morris Disability Questionnaire). Both mediator and outcome are continuous variables, and the outcome ranges between 0 and 1 with a higher score suggesting a severe pain-related disability condition. We adjust for the cluster size and the following individual-level baseline covariates: sex, age, diagnosis of 2 or more chronic medical conditions, anxiety and/or depression, number of different pain types, opioid dosage, Benzodiazepine receipt, and RMDQ and PEGS scores at baseline.

We estimate the cluster-average and individual-average natural indirect effects, natural direct effects, and spillover mediation effects with the estimators developed in Section 4; and model specifications follow Section 5. In Table 3, all methods provide similar point estimates for each mediation effect estimand. Specifically, the results suggest that the cognitive behavioral therapy intervention exerts negative nature indirect and direct effects, with 95% confidence intervals either excluding 0 or just crossing 0; an exception is the unstabilized machine learning estimator $\widehat{\mathrm{NIE}}_C^{\text{eif}_1\text{-ml-o}}$ whose interval appears much wider. In conclusion, a proportion of the protective effect from cognitive behavioral therapy on pain-related disability is likely transmitted through patient-reported pain impact. Furthermore, the estimated spillover mediation effects are generally close to zero and do not provide substantial evidence that the mediation is through unknown interactions with other same-cluster members.

7 Discussion

The primary contribution of this article is developing a semiparametric framework for assessing causal mediation with cluster-randomized experiments without ruling out informative cluster size and within-cluster interference. We formally establish the semiparametric efficiency theory for several mediation effect estimands and additional provide alternative strategies for operationalizing the natural (in)direct effect and spillover mediation effect estimators, where different representations of the nuisance functions are fitted with parametric working models or data-adaptive machine learners. Each estimator is doubly robust, with consistency guaranteed when one of two sets of nuisance models is correctly specified, but not necessary both. We additionally observe that the stabilization procedure in Section

Table 3: Mediation of the effect of a cognitive behavior therapy on pain-related disability by patient-reported pain impact, the Pain Program for Active Coping and Training study, 2014–2017. The numbers in parentheses are the associated 95% confidence interval.

Type	Estimator	NIE	NDE	SME	MP¶	SP¶
Cluster average	mf-par	−0.02 (−0.04,−0.01)	−0.04 (−0.07,−0.02)	0.00 (−0.02, 0.01)	33.7%	23.4%
	eif_1-par-o	−0.02 (−0.04,−0.01)	−0.03 (−0.07,−0.01)	−0.01 (−0.02, 0.01)	41.8%	29.2%
	eif_1-par-s	−0.03 (−0.06,−0.01)	−0.03 (−0.06, 0.00)	−0.01 (−0.03, 0.02)	51.3%	39.0%
	eif_1-ml-o	−0.05 (−0.11, 0.02)	−0.01 (−0.09, 0.06)	−0.02 (−0.06, 0.03)	77.3%	35.2%
	eif_1-ml-s	−0.03 (−0.05, 0.00)	−0.04 (−0.07, 0.00)	−0.01 (−0.04, 0.03)	41.8%	26.5%
	eif_2-par-o	−0.03 (−0.07, 0.00)	−0.03 (−0.07, 0.02)	−0.01 (−0.05, 0.03)	42.9%	38.4%
	eif_2-par-s	−0.02 (−0.05, 0.00)	−0.03 (−0.07, 0.00)	−0.01 (−0.03, 0.03)	41.6%	36.2%
	eif_2-ml-o	−0.02 (−0.06, 0.01)	−0.03 (−0.08, 0.01)	−0.03 (−0.11, 0.06)	41.7%	–
	eif_2-ml-s	−0.02 (−0.05, 0.01)	−0.04 (−0.08, 0.00)	−0.01 (−0.08, 0.05)	34.9%	71.3%
Individual average	mf-par	−0.02 (−0.03,−0.01)	−0.03 (−0.06,−0.01)	0.00 (−0.01, 0.01)	34.3%	23.0%
	eif_1-par-o	−0.02 (−0.04,−0.01)	−0.03 (−0.06, 0.00)	−0.01 (−0.02, 0.01)	37.5%	30.8%
	eif_1-par-s	−0.02 (−0.04,−0.01)	−0.03 (−0.06, 0.00)	−0.01 (−0.02, 0.01)	44.3%	37.8%
	eif_1-ml-o	−0.03 (−0.08, 0.01)	−0.02 (−0.07, 0.03)	−0.01 (−0.05, 0.03)	60.1%	30.8%
	eif_1-ml-s	−0.02 (−0.05, 0.00)	−0.03 (−0.07, 0.00)	−0.01 (−0.04, 0.02)	39.0%	27.6%
	eif_2-par-o	−0.02 (−0.06, 0.00)	−0.03 (−0.07, 0.02)	−0.01 (−0.05, 0.03)	41.0%	47.9%
	eif_2-par-s	−0.02 (−0.04, 0.01)	−0.03 (−0.07, 0.00)	−0.01 (−0.03, 0.02)	35.3%	42.2%
	eif_2-ml-o	−0.02 (−0.05, 0.02)	−0.04 (−0.08, 0.01)	−0.02 (−0.08, 0.05)	30.0%	–
	eif_2-ml-s	−0.01 (−0.04, 0.01)	−0.04 (−0.08, 0.00)	−0.01 (−0.07, 0.05)	25.9%	76.4%

¶ 'MP' is the mediation proportion, defined as the ratio between the NIE and the total effect (i.e., NIE+NDE). 'SP' is the spillover proportion, defined as the ratio between the SME and NIE. We only present the MP and SP when their values lies between 0 and 1.

4.4 can effectively improve the stability of the proposed estimator in finite samples, and will be recommended in practice.

Assumptions 3 and 5 are specific to the mediation context and are not directly verifiable from the observed data. Violation of Assumption 5 does not affect validity on estimation of the natural (in)direct effects, but estimators among the SME and IME only have interpretations under the interventional mediation analysis framework (Remark 1). If Assumption 3 fails to hold, the mediation effect estimators may not be causally interpretable, as expected from results for independent data Imai et al. (2010). In future work, we plan to develop a sensitivity analysis framework to assess vulnerability of our estimators under violation of Assumption 3. A viable strategy is to consider the semiparametric sensitivity analysis framework in Tchetgen and Shpitser (2012) by developing an interpretable sensitivity function to measure the degree of departure from Assumption 3. Then, bias-corrected estimators can be developed, and reporting the bias-corrected estimates based on a working sensitivity function with varying sensitivity parameters will provide insights into the tipping point for causal mediation results.

The intervention studied in a cluster-randomized experiment may have multiple components targeting more than one intermediate outcomes. Building upon the semiparametric efficiency theory we establish in this work, it would be relevant to extend the doubly robust and efficient causal mediation estimators to accommodate the scenario with two or more mediators. Depending on the scientific context and measurement structure, the mediators can be causally dependent or independent (Xia and Chan, 2022), leading to a different decomposition of cluster-average or individual-average treatment effect into an indirect effect and spillover effect estimand. When multiple mediators are of interest simultaneously in a cluster-randomized experiments, additional structural assumptions may be required to develop identification formulas and the efficient influence functions, which may pave the way for deriving efficient estimators to shed light on the more complex mediation mechanisms associated with multiple mediators.

Acknowledgement

Research in this article was supported by the Patient-Centered Outcomes Research Institute® (PCORI® Award ME-2023C1-31350). The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the views of PCORI®, its Board of Governors or Methodology Committee.

References

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006), "Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations." *Psychological methods*, 11, 142.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993), *Efficient and adaptive estimation for semiparametric models*, volume 4, Springer.

Charters, T. J., Kaufman, J. S., and Nandi, A. (2023), "A Causal Mediation Analysis for Investigating the Effect of a Randomized Cash-Transfer Program in Nicaragua," *American Journal of Epidemiology*, 192, 111–121.

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.

DeBar, L., Mayhew, M., Benes, L., Bonifay, A., Deyo, R. A., Elder, C. R., Keefe, F. J., Leo, M. C., McMullen, C., Owen-Smith, A., et al. (2022), "A primary care–based cognitive behavioral therapy intervention for long-term opioid users with chronic pain: a randomized pragmatic trial," *Annals of Internal Medicine*, 175, 46–55.

Field, C. A. and Welsh, A. H. (2007), "Bootstrapping clustered data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, 369–390.

Forastiere, L., Mealli, F., and VanderWeele, T. J. (2016), "Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using Bayesian principal stratification," *Journal of the American Statistical Association*, 111, 510–525.

Hudgens, M. G. and Halloran, M. E. (2008), "Toward causal inference with interference," *Journal of the American Statistical Association*, 103, 832–842.

Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Joe, H. (1997), *Multivariate models and multivariate dependence concepts*, CRC press.

Kahan, B. C., Li, F., Copas, A. J., and Harhay, M. O. (2023), "Estimands in cluster-randomized trials: choosing analyses that answer the right question," *International Journal of Epidemiology*, 52, 107–118.

Keane, M. P. (1994), "A computationally practical simulation estimator for panel data," *Econometrica: Journal of the Econometric Society*, 95–116.

Lee, H., Cashin, A. G., Lamb, S. E., Hopewell, S., Vansteelandt, S., VanderWeele, T. J., MacKinnon, D. P., Mansell, G., Collins, G. S., Golub, R. M., et al. (2021), "A guideline for reporting mediation analyses of randomized trials and observational studies: the AGReMA statement," *Jama*, 326, 1045–1056.

Masarotto, G. and Varin, C. (2012), "Gaussian copula marginal regression," *Electronic Journal of Statistics*, 6, 1517–1549.

— (2017), "Gaussian copula regression in R," *Journal of Statistical Software*, 77, 1–26.

Mi, X., Miwa, T., and Hothorn, T. (2009), "mvtnorm: New numerical algorithm for multivariate normal probabilities," *R Journal 1 (2009), Nr. 1*, 1, 37–39.

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

Moreno-Betancur, M. and Carlin, J. B. (2018), "Understanding interventional effects: a more natural approach to mediation analysis?" *Epidemiology*, 29, 614–617.

Murray, D. M. et al. (1998), *Design and Analysis of Group-Randomized Trials*, volume 29, Oxford University Press, USA.

Ogburn, E. L., Sofrygin, O., Diaz, I., and Van der Laan, M. J. (2022), "Causal inference for social network data," *Journal of the American Statistical Association*, 1–15.

Park, C. and Kang, H. (2023), "Assumption-lean analysis of cluster randomized trials in infectious diseases for intent-to-treat effects and network effects," *Journal of the American Statistical Association*, 118, 1195–1206.

Park, S. and Kaplan, D. (2015), "Bayesian causal mediation analysis for group randomized designs with homogeneous and heterogeneous effects: Simulation and case study," *Multivariate behavioral research*, 50, 316–333.

Phillips, R. V., Van Der Laan, M. J., Lee, H., and Gruber, S. (2023), "Practical considerations for specifying a super learner," *International Journal of Epidemiology*, 52, 1276–1285.

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

VanderWeele, T. J. (2010), "Direct and indirect effects for neighborhood-based clustered and longitudinal data," *Sociological methods & research*, 38, 515–544.

VanderWeele, T. J., Hong, G., Jones, S. M., and Brown, J. L. (2013), "Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention," *Journal of the American Statistical Association*, 108, 469–482.

VanderWeele, T. J. and Vansteelandt, S. (2009), "Conceptual issues concerning mediation, interventions and composition," *Statistics and its Interface*, 2, 457–468.

Wang, B., Park, C., Small, D. S., and Li, F. (2023), "Model-robust and efficient covariate adjustment for cluster-randomized experiments," *Journal of the American Statistical Association*, 1–28.

Williams, N. J. (2016), "Multilevel mechanisms of implementation strategies in mental health: integrating theory, research, and practice," *Administration and Policy in Mental Health and Mental Health Services Research*, 43, 783–798.

Xia, F. and Chan, K. C. G. (2022), "Decomposition, identification and multiply robust estimation of natural mediation effects with multiple mediators," *Biometrika*, 109, 1085–1100.

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009), "Testing multilevel mediation using hierarchical linear models: Problems and solutions," *Organizational Research Methods*, 12, 695–719.

Zhou, X. (2022), "Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 794–821.

Supplementary Material to "Semiparametric causal mediation analysis in cluster-randomized experiments"

Contents

A Supporting information on implementation of the proposed semiparametric estimators

A.1 On the use of a multivariate copula to model the joint mediator probability

This section provides technical details on the estimation of $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n)$ based on a multivariate copula $\mathcal{C}$, with a focus on the Gaussian copula (Masarotto and Varin, 2012). Recall that we parameterize $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n)$ into two variationally independent components: the marginal mediator probability for each individual $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ and a copula $\mathcal{C}$ characterizing the association structure across $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ for $j = 1, \dots, n$. Specifically, the Sklar's theorem (see, e.g., Joe (1997)) suggests that there exists a n-variate copula $\mathcal{C}$ such that

$$P_{\boldsymbol{M}|A,\boldsymbol{C},N}(\boldsymbol{m}|a, \boldsymbol{c}, n) = \mathcal{C}(\mathcal{K}_{\cdot 1}(m_{\cdot 1}), \cdots, \mathcal{K}_{\cdot n}(m_{\cdot n})|a, \boldsymbol{c}, n), \tag{s1}$$

where $\mathcal{K}_{\cdot j}(m_{\cdot j}) \equiv \mathcal{K}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n) := \int_{-\infty}^{m_{\cdot j}} \kappa_{\cdot j}(a, t, \boldsymbol{c}, n)\mathrm{d}t$ is the CDF of $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ and $\mathcal{C}(u_{\cdot 1}, \dots, u_{\cdot n}|a, \boldsymbol{c}, n)$ is a n-variate copula conditional on $\{A = a, \boldsymbol{C} = \boldsymbol{c}, N = n\}$ with support contained in $\{u_{\cdot 1}, \cdots, u_{\cdot n}\} \in [0, 1]^n$. When the mediator is continuous, $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n)$

and $\kappa_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n)$ can be re-expressed as

$$\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n) = c(\mathcal{K}_{\cdot 1}(m_{\cdot 1}), \cdots, \mathcal{K}_{\cdot n}(m_{\cdot n})|a, \boldsymbol{c}, n) \prod_{j=1}^{n} \kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n),$$

$$\kappa^c_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = \int \kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n) \mathrm{d}m_{\cdot j},$$

respectively, where $c(u_{\cdot 1}, \ldots, u_{\cdot n}|a, \boldsymbol{c}, n) = \frac{\partial^n}{\partial u_{\cdot 1} \cdots \partial u_{\cdot n}} \mathcal{C}(u_{\cdot 1}, \ldots, u_{\cdot n}|a, \boldsymbol{c}, n)$ is the density of the copula $\mathcal{C}$. When the mediator is discrete, expressions of $\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ and $\kappa^c_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n)$ includes multi-dimensional integrals. To simplify our discussion, we consider $M_{\cdot j} \in \{0, 1, \ldots, R\}$ (R can be $+\infty$) when the mediator is discrete. Then, with a discrete mediator, the expressions of $\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ and $\kappa^c_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n)$ are

$$\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n) = \int_{D_{\cdot 1}} \cdots \int_{D_{\cdot n}} c(u_{\cdot 1}, \cdots, u_{\cdot n}|a, \boldsymbol{c}, n) \mathrm{d}u_{\cdot 1} \ldots \mathrm{d}u_{\cdot n},$$

$$\kappa^c_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = \int_{\overline{D}_{\cdot 1}} \cdots \int_{\overline{D}_{\cdot n}} c(u_{\cdot 1}, \cdots, u_{\cdot n}|a, \boldsymbol{c}, n) \mathrm{d}u_{\cdot 1} \ldots \mathrm{d}u_{\cdot n},$$

where $D_{\cdot l} = [\mathcal{K}_{\cdot l}(m_{\cdot l} - 1), \mathcal{K}_{\cdot l}(m_{\cdot l})]$ for $l = 1, \ldots, n$, and $\overline{D}_{\cdot l} = D_{\cdot l}$ if $l \neq j$ and $\overline{D}_{\cdot l} = (-\infty, \infty)$ if $l = j$.

We use a Gaussian copula (5) to model the association structure, and below we provide the specific procedures on estimating the unknown parameters $\boldsymbol{\rho}$. First consider the scenario with a continuous mediator, where the joint probability of $\boldsymbol{M}$ conditional on A, $\boldsymbol{C}$, and N has the following explicit expression

$$\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n) = \boldsymbol{\phi}_n(\boldsymbol{\epsilon}; \boldsymbol{I}_n(\boldsymbol{\rho})) \prod_{j=1}^{n} \frac{\kappa_j(a, m_{\cdot j}, \boldsymbol{c}, n)}{\phi(\mathcal{K}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n))},$$

where $\boldsymbol{\epsilon} = [\epsilon_{\cdot 1}, \ldots, \epsilon_{\cdot n}]^T$ with $\epsilon_{\cdot j} = \Phi^{-1}(\mathcal{K}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n))$. This suggests that, as long as we have an appropriate estimator $\widehat{\kappa}_{\cdot j}$, $\widehat{\boldsymbol{\rho}}$ can be obtained by maximizing the following pseudo-likelihood which replaces $\kappa_{\cdot j}$ in the likelihood with its estimator:

$$L(\boldsymbol{\rho}) = \prod_{i=1}^{K} \left\{ \boldsymbol{\phi}_{N_i}\left(\widehat{\boldsymbol{\epsilon}}_i; \boldsymbol{I}_{N_i}(\boldsymbol{\rho})\right) \prod_{j=1}^{N_i} \frac{\widehat{\kappa}_{ij}(A_i, M_{ij}, \boldsymbol{C}_i, N_i)}{\phi(\widehat{\epsilon}_{ij})} \right\} \propto \prod_{i=1}^{K} \boldsymbol{\phi}_{N_i}\left(\widehat{\boldsymbol{\epsilon}}_i; \boldsymbol{I}_{N_i}(\boldsymbol{\rho})\right),$$

where $\widehat{\boldsymbol{\epsilon}}_i = [\epsilon_{i1}, \ldots, \epsilon_{iN_i}]^T$ with $\widehat{\epsilon}_{ij} = \Phi^{-1}(\widehat{\mathcal{K}}_{ij}(A_i, M_{ij}, \boldsymbol{C}_i, N_i))$. After obtaining $\widehat{\boldsymbol{\rho}}$, we can calculate point estimates of κ and $\kappa_{\cdot(-j)}$ based on

$$\widehat{\kappa}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \boldsymbol{\phi}_n\left(\widehat{\boldsymbol{\epsilon}}; \boldsymbol{I}_n(\widehat{\boldsymbol{\rho}})\right) \prod_{j=1}^{n} \frac{\widehat{\kappa}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)}{\phi(\widehat{\epsilon}_{\cdot j})},$$

$$\widehat{\kappa}_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = \boldsymbol{\phi}_{n-1}\left(\widehat{\boldsymbol{\epsilon}}_{\cdot(-j)}; \boldsymbol{I}_{n,(-j)}(\widehat{\boldsymbol{\rho}})\right) \prod_{l=1, l\neq j}^{n} \frac{\widehat{\kappa}_{\cdot l}(a, m_{\cdot l}, \boldsymbol{c}, n)}{\phi(\widehat{\epsilon}_{\cdot l})},$$

where $\widehat{\boldsymbol{\epsilon}}$ is the estimate of $\boldsymbol{\epsilon}$ based on $\widehat{\kappa}_{\cdot j}$, $\widehat{\boldsymbol{\epsilon}}_{\cdot(-j)}$ is $\widehat{\boldsymbol{\epsilon}}$ excluding its j-th element, and $\boldsymbol{I}_{n,(-j)}(\widehat{\boldsymbol{\rho}})$

is $\boldsymbol{I}_n(\widehat{\boldsymbol{\rho}})$ after dropping its j-th column and j-th row.

Next we consider the scenario with a discrete mediator. The discrete scenario is more complicated because the probability mass function $\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ does not have a closed-form representation. As shown in Masarotto and Varin (2012), $\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ has the following expression that contains a n-dimensional normal integral:

$$\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n) = \int_{E_{\cdot 1}} \cdots \int_{E_{\cdot n}} \boldsymbol{\phi}_n(\boldsymbol{\epsilon}; \boldsymbol{I}_n(\boldsymbol{\rho})) \mathrm{d}\epsilon_{\cdot 1} \dots \mathrm{d}\epsilon_{\cdot n},$$

where $E_{\cdot j} = [\Phi^{-1}(\mathcal{K}_{\cdot j}(a, m_{\cdot j} - 1, \boldsymbol{c}, n)), \Phi^{-1}(\mathcal{K}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n))]$. The n-dimensional normal integral in $\kappa^c(a, \boldsymbol{m}, \boldsymbol{c}, n)$ does not have an explicit expression, but a popular numerical method to calculate this integral is through the Geweke-Hajivassiliou-Keane (GHK) simulator (Keane, 1994), available in several **R** packages (e.g., the **gcmr** package (Masarotto and Varin, 2017) and **mvtnorm** package (Mi et al., 2009)). The pseudo-likelihood for $\boldsymbol{\rho}$ can be written as

$$L(\boldsymbol{\rho}) = \prod_{i=1}^{K} \int_{\widehat{E}_{i1}} \cdots \int_{\widehat{E}_{iN_i}} \boldsymbol{\phi}_n(\boldsymbol{\epsilon}; \boldsymbol{I}_n(\boldsymbol{\rho})) \mathrm{d}\epsilon_{i1} \dots \mathrm{d}\epsilon_{iN_i},$$

where $\widehat{E}_{ij} = \left[\Phi^{-1}(\widehat{\mathcal{K}}_{ij}(A_i, M_{ij} - 1, \boldsymbol{C}_i, N_i)), \Phi^{-1}(\widehat{\mathcal{K}}_{ij}(A_i, M_{ij}, \boldsymbol{C}_i, N_i))\right]$. Analogously to the continuous mediator scenario, $\widehat{\boldsymbol{\rho}}$ can be obtained by maximizing $L(\boldsymbol{\rho})$. After obtaining $\widehat{\boldsymbol{\rho}}$, we can calculate point estimates of κ and $\kappa_{\cdot(-j)}$ based on

$$\widehat{\kappa}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \int_{\widehat{E}_{\cdot 1}} \cdots \int_{\widehat{E}_{\cdot n}} \boldsymbol{\phi}_n(\boldsymbol{\epsilon}; \boldsymbol{I}_n(\boldsymbol{\rho})) \mathrm{d}\epsilon_{\cdot 1} \dots \mathrm{d}\epsilon_{\cdot n},$$
$$\widehat{\kappa}_{\cdot(-j)}(a, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n) = \int_{\widehat{\widehat{E}}_{\cdot 1}} \cdots \int_{\widehat{\widehat{E}}_{\cdot n}} \boldsymbol{\phi}_n(\boldsymbol{\epsilon}; \boldsymbol{I}_n(\boldsymbol{\rho})) \mathrm{d}\epsilon_{\cdot 1} \dots \mathrm{d}\epsilon_{\cdot n},$$

where $\widehat{\widehat{E}}_{\cdot l} = \widehat{E}_{\cdot l}$ if $l \neq j$ and $\widehat{\widehat{E}}_{\cdot l} = (-\infty, \infty)$ if $l = j$.

In what follows, we demonstrate that in the special case when (i) $\kappa_{\cdot j}(a, m, \boldsymbol{c}, n)$ is continuous and modeled by linear regression and (ii) $\mathcal{C}$ is modelled by a Gaussian copula with an exchangeable association structure, $\kappa(a, \boldsymbol{m}, \boldsymbol{c}, n)$ is precisely a linear mixed model and ρ is the intracluster correlation parameter (Murray et al., 1998). For illustration, we shall assume that $\kappa_{\cdot j}(a, m_{\cdot j}, \boldsymbol{c}, n)$ follows the following linear regression that is used in our simulation and application studies

$$M_{\cdot j} = \gamma_0 + \gamma_1 A + \gamma_2 A \times N + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \boldsymbol{X}_{\cdot j} + \gamma_5 N + \epsilon_{M_{\cdot j}}, \quad \text{(s2)}$$

where $\epsilon_{M_{\cdot j}} \sim N(0, \sigma_M^2)$. Notice that our results hold for more general linear regression settings by adding high-order terms of N, $\boldsymbol{V}$, and $\boldsymbol{X}_{\cdot j}$ and by adding arbitrary interaction terms among A, N, $\boldsymbol{M}$, and $\boldsymbol{X}_{\cdot j}$ in the mean structure of (s2). If we further assume $\mathcal{C}$ following a Gaussian copula with with an exchangeable association structure, then (s1) suggests that $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$ follows the following multivariate normal distribution

$$\boldsymbol{M}|A, \boldsymbol{C}, N \sim N\left(\boldsymbol{\mu}_{\boldsymbol{M}}(\boldsymbol{\gamma}), \boldsymbol{\Sigma}(\sigma_M^2, \rho)\right) \quad \text{(s3)}$$

where $\boldsymbol{\mu}_{\boldsymbol{M}}(\boldsymbol{\gamma}) = [\boldsymbol{\mu}_{M_{\cdot 1}}(\boldsymbol{\gamma}), \cdots, \boldsymbol{\mu}_{M_{\cdot n}}(\boldsymbol{\gamma})]^T$, $\boldsymbol{\mu}_{M_{\cdot 1}}(\boldsymbol{\gamma})$ is the mean structure of (s2), and

$\boldsymbol{\Sigma}(\sigma_M^2, \rho)$ is a n-by-n covariance matrix with diagonal elements σ_M^2 and all other elements $\rho\sigma_M^2$. This further implies that $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$ follows the following linear mixed-effects model

$$\boldsymbol{M} = \gamma_0 + \gamma_1 A + \gamma_2 A \times N + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{X}^\star \boldsymbol{\gamma}_4 + \gamma_5 N + b + \boldsymbol{\epsilon}_M^\star, \tag{s4}$$

where $\boldsymbol{X}^\star = [\boldsymbol{X}_{\cdot 1}^T, \ldots, \boldsymbol{X}_{\cdot N}^T]^T$, $\boldsymbol{\epsilon}_M^\star = [\epsilon_{M.1}^\star, \ldots, \epsilon_{M.N}^\star]^T$ is an individual-level random effect with $\epsilon_{M.j}^\star \overset{i.i.d.}{\sim} N(0, (1-\rho)\sigma_M^2)$, and $b \sim N(0, \rho\sigma_M^2)$ is a cluster-level random effect. Therefore, by the definition of intracluster correlation coefficient (ICC), we have

$$\text{ICC} = \frac{\text{Var}(b)}{\text{Var}(b) + \text{Var}(\epsilon_{M.j}^\star)} = \rho.$$

A.2 Supporting information on the stabilization procedures

We first briefly summarize the stabilization procedures for each semiparametric estimators. Procedures for obtaining stabilized $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a, a^*)$ (with $V \in \{I, C\}$) are presented in Section 4.4 in the main manuscript. Next, we summarize steps to obtain stabilized $\widehat{\tau}_V^{\text{eif}_1\text{-par}}$ (with $V \in \{I, C\}$):

1. Obtain $\widehat{h}_{nuisance}^{(1)} = \{\widehat{\eta}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot(-j)}^c\}$ based on the original procedure.
2. Define $\varrho(N) = \frac{1}{N}$ and 1 if $V = C$ and I respectively. For all clusters with treatment condition $A_i = 1$, fit a GLM for the conditional mean of Y_{ij}, using $\upsilon(\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i))$ as an offset term with $\varrho(N_i)\widehat{w}^{(2)}(1, 0, 1, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the only covariate (without an intercept), where $\upsilon(\cdot)$ is a canonical link function. Then, update $\widehat{\eta}_{\cdot j}^{\text{stab}}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \upsilon^{-1}(\upsilon(\widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)) + \widehat{\beta}\varrho(n)\widehat{w}^{(2)}(1, 0, 1, \boldsymbol{m}, \boldsymbol{c}, n))$, where $\widehat{\beta}$ is the estimated regression coefficient.
3. For $i = 1, \ldots, K$, calculate $\widehat{\psi}_\tau^{\text{stab}}(\boldsymbol{O}_i)$ by

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\Bigg\{\frac{A_i}{\pi}\left\{\widehat{u}_{ij}^{(2)}(1, 0, M_{ij}, \boldsymbol{C}_i, N_i) - \widehat{u}_{ij}^{(4)}(1, 1, 0, \boldsymbol{C}_i, N_i)\right\}$$
$$+ \frac{1-A}{1-\pi}\left\{\widehat{u}_{ij}^{(3)}(1, 1, \boldsymbol{M}_{i(-j)}, \boldsymbol{C}_i, N_i) - \widehat{u}_{ij}^{(4)}(1, 1, 0, \boldsymbol{C}_i, N_i)\right\} + \widehat{u}_{ij}^{(4)}(1, 1, 0, \boldsymbol{C}_i, N_i)\Bigg\}$$

   based on the updated estimates $\widehat{h}_{nuisance}^{(1)} = \{\widehat{\eta}_{\cdot j}^{\text{stab}}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot(-j)}^c\}$. Then, substitute $\widehat{\psi}_\tau^{\text{stab}}(\boldsymbol{O}_i)$ into (7) to obtain the stabilized $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a, a^*)$.

The stabilization procedures for $\widehat{\theta}_V^{\text{eif}_2\text{-par}}(a, a^*)$ and $\widehat{\tau}_V^{\text{eif}_2\text{-par}}$ (for both $V \in \{I, C\}$) share the same spirit to $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a, a^*)$ and $\widehat{\tau}_V^{\text{eif}_1\text{-par}}$, respectively. We therefore only narrate the procedures to obtain stabilized $\widehat{\theta}_V^{\text{eif}_2\text{-par}}(a, a^*)$, where stabilized $\widehat{\tau}_V^{\text{eif}_2\text{-par}}$ can be similarly obtained:

1. Obtain $\widehat{h}_{nuisance}^{(2)} = \{\widehat{\eta}_{\cdot j}, \widehat{\eta}_{\cdot j}^\star, \widehat{\eta}_{\cdot j}^\dagger, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot j}^\star, \widehat{s}\}$ based on the original procedure.
2. Define $\varrho(N) = \frac{1}{N}$ and 1 if $V = C$ and I respectively. For all clusters with treatment condition $A_i = a$, fit a GLM for the conditional mean of Y_{ij}, using $\upsilon(\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i))$ as an offset term with $\varrho(N_i)\widehat{w}^{(1)}(a, a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the only covariate (without an intercept), where $\upsilon(\cdot)$ is a canonical link function. Then, update $\widehat{\eta}_{\cdot j}^{\text{stab}}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \upsilon^{-1}(\upsilon(\widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)) + \widehat{\beta}\varrho(n)\widehat{w}^{(1)}(a, a^*, \boldsymbol{m}, \boldsymbol{c}, n))$, where $\widehat{\beta}$ is the estimated regression

coefficient.

3. For $i = 1, \dots, K$, calculate $\widehat{\psi}_\theta^{\text{stab}}(a, a^*; \boldsymbol{O}_i)$ by

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\left\{\frac{\mathbb{I}(A=a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left[\widehat{\eta}_{ij}^{\text{stab}}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i) - \widehat{u}_{ij}^{(1)}(a, a^*, \boldsymbol{C}_i, N_i)\right] + \widehat{u}_{ij}^{(1)}(a, a^*, \boldsymbol{C}_i, N_i)\right\}$$

based on the updated estimates $\widehat{h}_{nuisance}^{(2)} = \{\widehat{\eta}_{\cdot j}^{\text{stab}}, \widehat{\eta}_{\cdot j}^{\star}, \widehat{\eta}_{\cdot j}^{\dagger}, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot j}^{\star}, \widehat{s}\}$. Then, substitute $\widehat{\psi}_\theta^{\text{stab}}(a, a^*; \boldsymbol{O}_i)$ into (7) to obtain the stabilized $\widehat{\theta}_C^{\text{eif}_2\text{-par}}(a, a^*)$.

Following a similar strategy, we can stabilize the machine learning estimators $\widehat{\zeta}^{\text{eif}_1\text{-ml}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-ml}}$, with $\zeta \in \{\theta_C(a, a^*), \theta_I(a, a^*), \tau_C, \tau_I\}$, where the key difference is that now the original nuisance function estimator in Step 1 is obtained based on the cross-fitting procedure. As a concrete example, we present the detail procedure to obtain the stabilized $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}$, and other stabilized machine learning estimators can be obtained similarly:

1. Obtain the original estimator $\widehat{h}_{nuisance}^{(1)} = \{\widehat{\eta}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot(-j)}^c\}$ based on the standard cross-fitting procedure. Specifically, we evenly partition the data into R non-overlapping groups $\{\mathcal{S}_1, \dots, \mathcal{S}_R\}$. For each $r \in \{1, \dots, R\}$, the nuisance function estimates on group $\mathcal{S}_r$, $\widehat{h}_{nuisance}^{(1),r} = \{\widehat{\eta}_{\cdot j}^r, \widehat{\kappa}^{c,r}, \widehat{\kappa}_{\cdot j}^r, \widehat{\kappa}_{\cdot(-j)}^{c,r}\}$, is obtained by machine learners trained based on $\mathcal{S}_{-r}$. Therefore, the overall nuisance function estimates under the entire data set is the concatenation $\widehat{h}_{nuisance}^{(1)} = \{\widehat{h}_{nuisance}^{(1),1}, \dots, \widehat{h}_{nuisance}^{(1),R}\}$.
2. Define $\varrho(N) = \frac{1}{N}$ and 1 if $V = C$ and I respectively. For all clusters with treatment condition $A_i = a$, fit a GLM for the conditional mean of Y_{ij}, using $\upsilon(\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i))$ as an offset term with $\varrho(N_i)\widehat{w}^{(1)}(a, a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the only covariate (without an intercept), where $\upsilon(\cdot)$ is a canonical link function. Then, update $\widehat{\eta}_{\cdot j}^{\text{stab}}(a, \boldsymbol{m}, \boldsymbol{c}, n) = \upsilon^{-1}(\upsilon(\widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)) + \widehat{\beta}\varrho(n)\widehat{w}^{(1)}(a, a^*, \boldsymbol{m}, \boldsymbol{c}, n))$, where $\widehat{\beta}$ is the estimated regression coefficient.
3. For $i = 1, \dots, K$, calculate $\widehat{\psi}_\theta^{\text{stab}}(a, a^*; \boldsymbol{O}_i)$ by

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\left\{\frac{\mathbb{I}(A=a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left[\widehat{\eta}_{ij}^{\text{stab}}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i) - \widehat{u}_{ij}^{(1)}(a, a^*, \boldsymbol{C}_i, N_i)\right] + \widehat{u}_{ij}^{(1)}(a, a^*, \boldsymbol{C}_i, N_i)\right\}$$

based on the updated estimates $\widehat{h}_{nuisance}^{(1)} = \{\widehat{\eta}_{\cdot j}^{\text{stab}}, \widehat{\kappa}^c, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}_{\cdot(-j)}^c\}$. Then, substitute $\widehat{\psi}_\theta^{\text{stab}}(a, a^*; \boldsymbol{O}_i)$ into (7) to obtain the stabilized $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a, a^*)$.

A.3 Practical strategies on specification of the parametric working models

We provide specification strategies for the parametric working models of $h_{nuisance}^{(1)}$ and $h_{nuisance}^{(2)}$ with a focus on the scenario with a continuous mediator and outcome, which also aligns with the settings in our simulation study and applications. An extension to a binary/count outcome will be considered later. To model $h_{nuisance}^{(1)}$, we need specifying parametric working models on $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, where $\mathcal{C}$ is modeled by a Gaussian copula with an exchangeable structure and $\rho \in (-1, 1)$ is the only correlation parameter characterizing all off-diagonal elements in $\boldsymbol{I}_n(\rho)$. To model $\kappa_{\cdot j}$, we shall assume $M_{\cdot j}|\{A, \boldsymbol{C}, N\}$ following the linear regression (s2). Also, we assume that the outcome mean (i.e., $\eta_{\cdot j}$) follows the linear regression:

$$\mathbb{E}[Y_{\cdot j}|A=a,\boldsymbol{M},\boldsymbol{C},N]=\beta_{0,a}+\beta_{1,a}M_{\cdot j}+\beta_{2,a}\overline{M}_{\cdot(-j)}+\boldsymbol{\beta}_{3,a}^T\boldsymbol{V}+\boldsymbol{\beta}_{4,a}^T\boldsymbol{X}_{\cdot j}+\beta_{5,a}N, \qquad \text{(s5)}$$

where $\boldsymbol{\beta}_a=[\beta_{0,a},\beta_{1,a},\beta_{2,a},\boldsymbol{\beta}_{3,a}^T,\boldsymbol{\beta}_{4,a}^T,\beta_{5,a}]^T$ are treatment-specific unknown coefficients. Estimation of the unknown parameters in $\kappa_{\cdot j}$ and $\eta_{\cdot j}$ follows the standard regression theory:

1. Observing that $\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)=\mathbb{E}[Y_{\cdot j}|A=a,\boldsymbol{M},\boldsymbol{C},N]$ follows (s5), $\boldsymbol{\beta}_a$ can be estimated by running a linear regression of Y_{ij} on $\{M_{ij},\overline{M}_{i(-j)},\boldsymbol{V}_i,\boldsymbol{X}_{ij},N_i\}$ based on all clusters with treatment $A_i=a$.

2. Observing that $\kappa_{\cdot j}(A,M_{\cdot j},\boldsymbol{C},N)=f(M_{\cdot j}|A,\boldsymbol{C},N)$ follows (s2), the unknown coefficients $\boldsymbol{\gamma}=[\gamma_0,\gamma_1,\gamma_2,\boldsymbol{\gamma}_3^T,\boldsymbol{\gamma}_4^T,\gamma_5]^T$ can be estimated by running a linear regression of M_{ij} on $\{A_i,A_iN_i,\boldsymbol{V},\boldsymbol{X}_{ij},N_i\}$ and σ_M^2 can be estimated by the sample variance of the residuals.

As discussed in the main manuscript, the six nuisance functions $\{\eta_{\cdot j},\eta^\star_{\cdot j},\eta^\dagger_{\cdot j},\kappa_{\cdot j},\kappa^\star_{\cdot j},s\}$ in $h^{(2)}_{nuisance}$ are not variationally independent, therefore it is necessary to ensure compatibility among these working models. In what follows, we provide a set of compatible parametric working models for $h^{(2)}_{nuisance}$, which are also compatiable with the aforementioned working models for $h^{(1)}_{nuisance}$. Specifications of $\kappa_{\cdot j}$ and $\eta_{\cdot j}$ are identical to these introduced previously, and we only narrate specifications on $\{\eta^\star_{\cdot j},\eta^\dagger_{\cdot j},\kappa^\star_{\cdot j},s\}$:

1. $\eta^\star_{\cdot j}(a,a^*,\boldsymbol{C},N)=\zeta_{0,(a,a^*)}+\boldsymbol{\zeta}_{1,(a,a^*)}^T\boldsymbol{V}+\boldsymbol{\zeta}_{2,(a,a^*)}^T\boldsymbol{X}_{\cdot j}+\boldsymbol{\zeta}_{3,(a,a^*)}^T\overline{\boldsymbol{X}}_{\cdot(-j)}+\zeta_{4,(a,a^*)}N$, where $\boldsymbol{\zeta}_{(a,a^*)}=[\zeta_{0,(a,a^*)},\boldsymbol{\zeta}_{1,(a,a^*)}^T,\boldsymbol{\zeta}_{2,(a,a^*)}^T,\boldsymbol{\zeta}_{3,(a,a^*)}^T,\zeta_{4,(a,a^*)}]^T$ can be estimated by running a linear regression of $\widehat{\eta}_{ij}(a,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)$ on $\{\boldsymbol{V}_i,\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_{i(-j)},N_i\}$ based on all clusters with treatment $A=a^*$.

2. $\eta^\dagger_{\cdot j}(a,a^*,M_{\cdot j},\boldsymbol{C},N)=\psi_{0,(a,a^*)}+\psi_{1,(a,a^*)}M_{\cdot j}+\boldsymbol{\psi}_{2,(a,a^*)}^T\boldsymbol{V}+\boldsymbol{\psi}_{3,(a,a^*)}^T\boldsymbol{X}_{\cdot j}+\boldsymbol{\psi}_{4,(a,a^*)}^T\overline{\boldsymbol{X}}_{\cdot(-j)}+\psi_{5,(a,a^*)}N$, where $\boldsymbol{\psi}_{(a,a^*)}=[\psi_{0,(a,a^*)},\psi_{1,(a,a^*)},\boldsymbol{\psi}_{2,(a,a^*)}^T,\boldsymbol{\psi}_{3,(a,a^*)}^T,\boldsymbol{\psi}_{4,(a,a^*)}^T,\psi_{5,(a,a^*)}]^T$ can be estimated by running a linear regression of $\frac{\widehat{\eta}_{ij}(a,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)\widehat{\kappa}_{ij}(a,M_{ij},\boldsymbol{C}_i,N_i)}{\widehat{\kappa}^\star_{ij}(a^*,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)}$ on $\{M_{ij},\boldsymbol{V}_i,\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_{i(-j)},N_i\}$ based on all clusters with treatment $A=a^*$.

3. $\kappa^\star_{\cdot j}(A,\boldsymbol{M},\boldsymbol{C},N)=f(M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)$ follows a normal distribution with mean $\mathbb{E}\left\{M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N\right\}=\sum_n\mathbb{I}(N=n)\Big\{\delta_{0,n}+\delta_{1,n}A+\delta_{2,n}\overline{M}_{\cdot(-j)}+\boldsymbol{\delta}_{3,n}^T\boldsymbol{V}+\boldsymbol{\delta}_{4,n}^T\boldsymbol{X}_{\cdot j}+\boldsymbol{\delta}_{5,n}^T\overline{\boldsymbol{X}}_{\cdot(-j)}\Big\}$ and $\mathrm{Var}(M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)\approx\sigma_M^{\star 2}$ (approximation holds if either N is large or ρ is small). Here, the regression coefficients $\boldsymbol{\delta}_n=[\delta_{0,n},\delta_{1,n},\delta_{2,n},\boldsymbol{\delta}_{3,n}^T,\boldsymbol{\delta}_{4,n}^T,\boldsymbol{\delta}_{5,n}^T]^T$ can be estimated by running a linear regression of M_{ij} on $\{A_i,\overline{M}_{i(-j)},\boldsymbol{V}_i,\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_{i(-j)}\}$ based on clusters with cluster size $N=n$. Because the variance $\sigma_M^{\star 2}$ is a constant, $\sigma_M^{\star 2}$ can be estimated by using the empirical variance of residuals among individuals in all clusters.

4. $s(A,\boldsymbol{M},\boldsymbol{C},N)=f(A|\boldsymbol{M},\boldsymbol{C},N)$ follows a logistic regression such that $f(A=1|\boldsymbol{M},\boldsymbol{C},N)=\mathrm{expit}\left\{\sum_n\mathbb{I}(N=n)\left(\alpha_{0,n}+\alpha_{1,n}\overline{M}+\boldsymbol{\alpha}_{2,n}^T\boldsymbol{V}+\boldsymbol{\alpha}_{3,n}^T\overline{\boldsymbol{X}}\right\}\right)$, where $\boldsymbol{\alpha}_n=[\alpha_{0,n},\alpha_{1,n},\boldsymbol{\alpha}_{2,n}^T,\boldsymbol{\alpha}_{3,n}^T]^T$ can be obtained by running a logistic regression of A_i on $\{\overline{M}_i,\boldsymbol{V}_i,\overline{\boldsymbol{X}}_i\}$ based on all clusters with cluster size $N=n$.

Proof for compatibility is presented at the end of this section. Notice that the working models of $\{\kappa^\star_{\cdot j}, s\}$ require estimation of cluster-size specific coefficients, which may be practically infeasible to estimate when the number of clusters (K) is small. More parsimonious parametric models of $\{\kappa^\star_{\cdot j}, s\}$ can be achieved by adding interaction terms between N and the other predictors in the parametric models. For example, one can consider the following model for the mean structure of $\kappa^\star_{\cdot j}$:

$$\begin{aligned}\mathbb{E}[M_{\cdot j}|A, \boldsymbol{M}_{\cdot(-j)}, \boldsymbol{C}, N] =& \delta_0 + \delta_1 A + \delta_2 \overline{M}_{\cdot(-j)} + \boldsymbol{\delta}_3^T \boldsymbol{V} + \boldsymbol{\delta}_4^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\delta}_5^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \\ & N \times \{\delta_6 + \delta_7 A + \delta_8 \overline{M}_{\cdot(-j)} + \boldsymbol{\delta}_9^T \boldsymbol{V} + \boldsymbol{\delta}_{10}^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\delta}_{11}^T \overline{\boldsymbol{X}}_{\cdot(-j)}\}, \quad \text{(s6)}\end{aligned}$$

where the second row includes first-order interaction terms between the cluster size N and other predictors $\{A_i, \overline{M}_{i(-j)}, \boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}\}$. One can also add a second-order interaction term between N^2 and $\{A_i, \overline{M}_{i(-j)}, \boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}\}$ to have a better approximation to the parametric model in Point 3. Similarly, the logistic model for $s(A, \boldsymbol{M}, \boldsymbol{C}, N)$ can be approximated by

$$f(A=1|\boldsymbol{M}, \boldsymbol{C}, N) = \text{expit}\left\{\alpha_0 + \alpha_1 \overline{M} + \boldsymbol{\alpha}_2^T \boldsymbol{V} + \boldsymbol{\alpha}_3^T \overline{\boldsymbol{X}} + N \times (\alpha_4 + \alpha_5 \overline{M} + \boldsymbol{\alpha}_6^T \boldsymbol{V} + \boldsymbol{\alpha}_7^T \overline{\boldsymbol{X}})\right\}. \quad \text{(s7)}$$

Next we consider the scenario with a binary/count outcome. In this scenario, the working models for $\{\kappa_{\cdot j}, \mathcal{C}, \kappa^\star_{\cdot j}, s\}$ remain the same but the models for $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}\}$ should be revised to reflect the change of outcome. We are interested in using the following generalized linear model with a log-link function to model $\eta_{\cdot j}(A, \boldsymbol{M}, \boldsymbol{C}, N)$:

$$\log\{\mathbb{E}[Y_{\cdot j}|A=a, \boldsymbol{M}, \boldsymbol{C}, N]\} = \beta_{0,a} + \beta_{1,a} M_{\cdot j} + \beta_{2,a} \overline{M}_{\cdot(-j)} + \boldsymbol{\beta}_{3,a}^T \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{5,a} N, \quad \text{(s8)}$$

where $\boldsymbol{\beta}_a = [\beta_{0,a}, \beta_{1,a}, \beta_{2,a}, \boldsymbol{\beta}_{3,a}^T, \boldsymbol{\beta}_{4,a}^T, \beta_{5,a}]^T$ are treatment-specific unknown coefficients. This model can be seen as a Poisson regression under a count outcome and a log-binomial regression under a binary outcome. The coefficients $\boldsymbol{\beta}_a$ can be estimated by fitting a quasi-Poisson regression based on a log-link function with Y_{ij} as response and $\{M_{ij}, \overline{M}_{i(-j)}, \boldsymbol{V}_i, \boldsymbol{X}_{\cdot j}, N\}$ as predictors, based on all clusters with treatment $A = a$. Specifications of $\{\eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}\}$ are given below:

1. $\log\left\{\eta^\star_{\cdot j}(a, a^*, \boldsymbol{C}, N)\right\} = \zeta_{0,(a,a^*)} + \boldsymbol{\zeta}_{1,(a,a^*)}^T \boldsymbol{V} + \boldsymbol{\zeta}_{2,(a,a^*)}^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\zeta}_{3,(a,a^*)}^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \zeta_{4,(a,a^*)} N + \zeta_{5,(a,a^*)} \frac{1}{N-1}$, where $\boldsymbol{\zeta}_{(a,a^*)} = [\zeta_{0,(a,a^*)}, \boldsymbol{\zeta}_{1,(a,a^*)}^T, \boldsymbol{\zeta}_{2,(a,a^*)}^T, \boldsymbol{\zeta}_{3,(a,a^*)}^T, \zeta_{4,(a,a^*)}, \zeta_{5,(a,a^*)}]^T$ can be estimated by running a quasi-Poisson regression based on a log-link function with $\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the response and $\{\boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}, N_i, \frac{1}{N_i-1}\}$ as predictors, based on all clusters with treatment $A = a^*$.

2. $\log\left\{\eta^\dagger_{\cdot j}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N)\right\} = \psi_{0,(a,a^*)} + \psi_{1,(a,a^*)} M_{\cdot j} + \boldsymbol{\psi}_{2,(a,a^*)}^T \boldsymbol{V} + \boldsymbol{\psi}_{3,(a,a^*)}^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\psi}_{4,(a,a^*)}^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \psi_{5,(a,a^*)} N + \psi_{6,(a,a^*)} \frac{1}{N-1}$, where $\boldsymbol{\psi}_{(a,a^*)} = [\psi_{0,(a,a^*)}, \psi_{1,(a,a^*)}, \boldsymbol{\psi}_{2,(a,a^*)}^T, \boldsymbol{\psi}_{3,(a,a^*)}^T, \boldsymbol{\psi}_{4,(a,a^*)}^T, \psi_{5,(a,a^*)}, \psi_{6,(a,a^*)}]^T$ can be estimated by running quasi-Poisson regression based on a log-link function with $\frac{\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)\widehat{\kappa}_{ij}(a, M_{ij}, \boldsymbol{C}_i, N_i)}{\widehat{\kappa}^\star_{ij}(a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)}$ as response and $\{M_{ij}, \boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}, N_i, \frac{1}{N_i-1}\}$ as predictors, based on all clusters with treatment $A = a^*$.

Proof of compatiability among the working models. We first consider the scenario with a continuous outcome. Regarding the data generation process of the observed variables, we assume $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$ following the multivariate normal distribution (s3) and $E[Y_{\cdot j}|A, \boldsymbol{M}, \boldsymbol{C}, N]$

following the linear model (s5). By construction, the parametric working models of $\eta_{\cdot j}$ and $\kappa_{\cdot j}$ hold. Next, we show that $\{\eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ align with the aforementioned data generation process, so they are mutually compatible.

1. Parametric model for $s(a, \boldsymbol{m}, \boldsymbol{c}, n)$

The population density function of $\boldsymbol{M}|\{A = a, \boldsymbol{C}, N\}$ is

$$f(\boldsymbol{M}|A = a, \boldsymbol{C}, N) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{M} - \boldsymbol{\mu}_a)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{M} - \boldsymbol{\mu}_a)\right\},$$

where $\boldsymbol{\mu}_a = [\mu_{a,1}, \ldots, \mu_{a,N}]^T$ is a N-by-1 vector with the j-th element $\mu_{a,j} = \gamma_0 + (\gamma_1 + \gamma_2 N)a + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \boldsymbol{X}_{\cdot j} + \gamma_5 N$, and $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\sigma_M^2, \rho)$ is the variance-covariance matrix. Notice that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = (\gamma_1 + \gamma_2 N)\boldsymbol{e}$, where $\boldsymbol{e}$ is a N-by-1 vector with all elements equal to 1. Therefore, we have that

$$\frac{f(\boldsymbol{M}|A = 0, \boldsymbol{C}, N)}{f(\boldsymbol{M}|A = 1, \boldsymbol{C}, N)} = \exp\left\{-\frac{1}{2}(\gamma_1 + \gamma_2 N)\boldsymbol{e}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{M} - \boldsymbol{\mu}_0) + \frac{1}{2}(\gamma_1 + \gamma_2 N)^2 \boldsymbol{e}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}\right\}.$$

Notice that $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_M^2}\left\{\frac{1}{1-\rho}\boldsymbol{E} - \frac{\rho}{(1-\rho)^2 + \rho(1-\rho)N}\boldsymbol{e}\boldsymbol{e}^T\right\}$ by the Sherman-Morrison formula, where $\boldsymbol{E}$ is a N-by-N identity matrix. We then have

$$\frac{f(\boldsymbol{M}|A = 0, \boldsymbol{C}, N)}{f(\boldsymbol{M}|A = 1, \boldsymbol{C}, N)} = \exp\left\{-\frac{\gamma_1 + \gamma_2 N}{2\sigma_M^2} \frac{N}{1 + \rho(N-1)} \left(\overline{M} - \gamma_0 - \gamma_1 - \gamma_2 N - \boldsymbol{\gamma}_3^T \boldsymbol{V} - \boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}} - \gamma_5 N\right)\right\}.$$

This implies that

$$\begin{aligned}
&f(A = 1|\boldsymbol{M}, \boldsymbol{C}, N) = \frac{f(\boldsymbol{M}|A = 1, \boldsymbol{C}, N) f(A = 1)}{\sum_{a=0}^{1} f(\boldsymbol{M}|A = a, \boldsymbol{C}, N) f(A = a)} \\
=&\text{expit}\left\{-\log\left(\frac{1-\pi}{\pi}\right) + \frac{\gamma_1 + \gamma_2 N}{2\sigma_M^2} \frac{N}{1 + \rho(N-1)} \left(\overline{M} - \gamma_0 - \gamma_1 - \gamma_2 N - \boldsymbol{\gamma}_3^T \boldsymbol{V} - \boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}} - \gamma_5 N\right)\right\}.
\end{aligned}$$

where $\text{expit}(x) = \frac{1}{1+\exp(-x)}$. This suggests that $f(A = 1|\boldsymbol{M}, \boldsymbol{C}, N)$ can be consistently estimated by using the logistic regression $f(A = 1|\boldsymbol{M}, \boldsymbol{C}, N) = \text{expit}\Big\{\sum_n \mathbb{I}(N = n)\Big[\alpha_{0,n} + \alpha_{1,n}\overline{M} + \boldsymbol{\alpha}_{2,n}^T \boldsymbol{V} + \boldsymbol{\alpha}_{3,n}^T \overline{\boldsymbol{X}}\Big]\Big\}$.

2. Parametric model for $\kappa^{\star}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n)$

Define $\boldsymbol{\Sigma}_{(-j)}$ as a $(N-1)$-by-$(N-1)$ matrix by extracting the j-th row and the j-th column from $\boldsymbol{\Sigma}$. Also, let $\boldsymbol{e}_{(-j)}$ be a $(N-1)$-by-1 vector with all elements 1. Based on the joint distribution of $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$, we have that $M_{\cdot j}|\{A, \boldsymbol{M}_{\cdot(-j)}, \boldsymbol{C}, N\}$ follows a normal distribution with mean

$$\begin{aligned}
&\mathbb{E}[M_{\cdot j}|A, \boldsymbol{M}_{\cdot(-j)}, \boldsymbol{C}, N] \\
=&\mathbb{E}[M_{\cdot j}|A, \boldsymbol{C}, N] + \rho\sigma_M^2 \boldsymbol{e}_{(-j)}^T \boldsymbol{\Sigma}_{(-j)}^{-1}\left\{\boldsymbol{M}_{\cdot(-j)} - \mathbb{E}[\boldsymbol{M}_{\cdot(-j)}|A, \boldsymbol{C}, N]\right\} \\
=&\gamma_0 + \gamma_1 A + \gamma_2 AN + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \gamma_5 N
\end{aligned}$$

$$+\frac{N-1}{1+\rho(N-2)}\left\{\overline{M}_{\cdot(-j)}-\gamma_0-\gamma_1 A-\gamma_2 AN-\boldsymbol{\gamma}_3^T\boldsymbol{V}-\boldsymbol{\gamma}_4^T\overline{\boldsymbol{X}}_{\cdot(-j)}-\gamma_5 N\right\}$$

and variance

$$\begin{aligned}\mathrm{Var}(M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)&=\sigma_M^2-\rho\sigma_M^2\boldsymbol{e}_{(-j)}^T\boldsymbol{\Sigma}_{(-j)}^{-1}\boldsymbol{e}_{(-j)}\rho\sigma_M^2\\&=\sigma_M^2(1-\rho)\left\{1+\frac{\rho}{1+\rho(N-2)}\right\}\\&\approx\sigma_M^2(1-\rho),\end{aligned}$$

where the approximation hold if either N is large or ρ is small. This suggests that $\mathbb{E}[M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N]$ can be consistently estimated by fitting a linear regression

$$\mathbb{E}[M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N]=\sum_n\mathbb{I}(N=n)\Big\{\delta_{0,n}+\delta_{1,n}A+\delta_{2,n}\overline{M}_{\cdot(-j)}+\boldsymbol{\delta}_{3,n}^T\boldsymbol{V}+\boldsymbol{\delta}_{4,n}^T\boldsymbol{X}_{\cdot j}+\boldsymbol{\delta}_{5,n}^T\overline{\boldsymbol{X}}_{\cdot(-j)}\Big\}.$$

Since $\mathrm{Var}(M_{\cdot j}|A,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)$ is approximately homoskedastic, one can consider using the empirical variance of the residuals $Y_{ij}-\widehat{\mathbb{E}}[M_{ij}|A_i,\boldsymbol{M}_{i(-j)},\boldsymbol{C}_i,N_i]$ as an estimator.

3. Parametric models for $\eta^\star_{\cdot j}(a,a^*,\boldsymbol{c},n)$

By definition, we have

$$\begin{aligned}&\eta^\star_{\cdot j}(a,a^*,\boldsymbol{C},N)=\mathbb{E}\left[\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\Big|A=a^*,\boldsymbol{C},N\right]\\&=\mathbb{E}\left[\beta_{0,a}+\beta_{1,a}M_{\cdot j}+\beta_{2,a}\overline{M}_{\cdot(-j)}+\boldsymbol{\beta}_{3,a}^T\boldsymbol{V}+\boldsymbol{\beta}_{4,a}^T\boldsymbol{X}_{\cdot j}+\beta_{5,a}N\Big|A=a^*,\boldsymbol{C},N\right]\\&=\beta_{0,a}+\boldsymbol{\beta}_{3,a}^T\boldsymbol{V}+\boldsymbol{\beta}_{4,a}^T\boldsymbol{X}_{\cdot j}+\beta_{5,a}N+\beta_{1,a}\mathbb{E}[M_{\cdot j}|A=a^*,\boldsymbol{C},N]+\beta_{2,a}\frac{1}{N-1}\sum_{l\neq j}\mathbb{E}[M_{\cdot l}|A=a^*,\boldsymbol{C},N],\end{aligned}$$

where $\mathbb{E}[M_{\cdot j}|A=a^*,\boldsymbol{C},N]=\gamma_0+\gamma_1a^*+\gamma_2a^*N+\boldsymbol{\gamma}_3^T\boldsymbol{V}+\boldsymbol{\gamma}_4^T\boldsymbol{X}_{\cdot j}+\gamma_5N$. This suggests that

$$\begin{aligned}\eta^\star_{\cdot j}(a,a^*,\boldsymbol{C},N)=&\beta_{0,a}+\beta_{1,a}\gamma_0+\beta_{2,a}\gamma_0+(\beta_{1,a}\gamma_1+\beta_{2,a}\gamma_1)a^*+(\boldsymbol{\beta}_{3,a}+\beta_{1,a}\gamma_3+\beta_{2,a}\gamma_3)^T\boldsymbol{V}\\&+(\boldsymbol{\beta}_{4,a}+\beta_{1,a}\boldsymbol{\gamma}_4)^T\boldsymbol{X}_{\cdot j}+\beta_{2,a}\boldsymbol{\gamma}_4^T\overline{\boldsymbol{X}}_{\cdot(-j)}+(\beta_{5,a}+(\beta_{1,a}+\beta_{2,a})(\gamma_2a^*+\gamma_5))N.\end{aligned}$$

Therefore, one can consider the working linear regression $\eta^\star_{\cdot j}(a,a^*,\boldsymbol{C},N)=\zeta_{0,(a,a^*)}+\boldsymbol{\zeta}_{1,(a,a^*)}^T\boldsymbol{V}+\boldsymbol{\zeta}_{2,(a,a^*)}^T\boldsymbol{X}_{\cdot j}+\boldsymbol{\zeta}_{3,(a,a^*)}^T\overline{\boldsymbol{X}}_{\cdot(-j)}+\zeta_{4,(a,a^*)}N$, where unknown coefficients can be estimated by running a linear regression of $\widehat{\eta}_{ij}(a,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)$ on $\{\boldsymbol{V}_i,\boldsymbol{X}_{ij},\overline{\boldsymbol{X}}_{i(-j)},N_i\}$ based on all clusters with treatment $A=a^*$.

4. Parametric models for $\eta^\dagger_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$.

By the definition of $\eta^\dagger_{\cdot j}$, we have that

$$\eta^\dagger_{\cdot j}(a,a^*,M_{\cdot j},\boldsymbol{C},N)=\int\eta_{\cdot j}(a,M_{\cdot j},\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)f(\boldsymbol{m}_{\cdot(-j)}|A=a^*,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}_{\cdot(-j)}$$

$$= \int \left\{ \beta_{0,a} + \beta_{1,a} M_{\cdot j} + \frac{\beta_{2,a}}{N-1} \sum_{l \neq j} m_{\cdot l} + \boldsymbol{\beta}_{3,a}^T \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{5,a} N \right\} f(\boldsymbol{m}_{\cdot(-j)} | A = a^*, \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}_{\cdot(-j)}$$
$$= \beta_{0,a} + \beta_{1,a} M_{\cdot j} + \boldsymbol{\beta}_{3,a}^T \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{5,a} N + \beta_{2,a} \left\{ \gamma_0 + \gamma_1 a^* + \gamma_2 a^* N + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \gamma_5 N \right\}$$
$$= \beta_{0,a} + \beta_{2,a}(\gamma_0 + \gamma_1 a^*) + \beta_{1,a} M_{\cdot j} + (\boldsymbol{\beta}_{3,a}^T + \beta_{2,a} \boldsymbol{\gamma}_3^T) \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{2,a} \boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}}_{\cdot(-j)}$$
$$+ (\beta_{5,a} + \beta_{2,a}\gamma_2 a^* + \beta_{2,a}\gamma_5) N$$

Therefore, one can consider the working linear regression $\eta_{\cdot j}^{\dagger}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N) = \psi_{0,(a,a^*)} + \psi_{1,(a,a^*)} M_{\cdot j} + \boldsymbol{\psi}_{2,(a,a^*)}^T \boldsymbol{V} + \boldsymbol{\psi}_{3,(a,a^*)}^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\psi}_{4,(a,a^*)}^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \psi_{5,(a,a^*)} N$, where unknown coefficients can be estimated by running a linear regression of $\frac{\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i) \widehat{\kappa}_{ij}(a, M_{ij}, \boldsymbol{C}_i, N_i)}{\widehat{\kappa}_{ij}^{\star}(a^*, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)}$ on $\{M_{ij}, \boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}, N_i\}$ based on all clusters with treatment $A = a^*$.

We next consider the scenario with a binary/count outcome, where we still assume $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$ following the multivariate normal distribution (s3), but now $E[Y_{\cdot j}|A, \boldsymbol{M}, \boldsymbol{C}, N]$ follows the generalized linear model (s8). The parametric models for $\{\kappa_{\cdot j}, \mathcal{C}, \kappa_{\cdot j}^{\star}, s\}$ remain the same to these in the scenario with a continuous outcome, next we only elucidate the parametric working models for $\{\eta_{\cdot j}^{\star}, \eta_{\cdot j}^{\dagger}\}$:

1^*. Parametric models for $\eta_{\cdot j}^{\star}(a, a^*, \boldsymbol{c}, n)$ with a binary/count outcome

Based on the joint distribution of $\boldsymbol{M}|\{A, \boldsymbol{C}, N\}$, we have

$$\begin{bmatrix} M_{\cdot j} \\ \overline{M}_{\cdot(-j)} \end{bmatrix} \Big| \{A = a^*, \boldsymbol{C}, N\} \sim N \left(\begin{bmatrix} \gamma_0 + \gamma_1 a^* + \gamma_2 a^* N + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \boldsymbol{X}_{\cdot j} + \gamma_5 N \\ \gamma_0 + \gamma_1 a^* + \gamma_2 a^* N + \boldsymbol{\gamma}_3^T \boldsymbol{V} + \boldsymbol{\gamma}_4^T \boldsymbol{X}_{\cdot j} + \gamma_5 N \end{bmatrix}, \begin{bmatrix} \sigma_M^2 & \rho\sigma_M^2 \\ \rho\sigma_M^2 & \frac{1+(N-2)\rho}{N-1}\sigma_M^2 \end{bmatrix} \right)$$

Therefore, we have that

$$\begin{aligned}
\eta_{\cdot j}^{\star}(a, a^*, \boldsymbol{C}, N) =& \mathbb{E}\left[\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) \Big| A = a^*, \boldsymbol{C}, N \right] \\
=& \mathbb{E}\left[e^{\beta_{0,a} + \beta_{1,a} M_{\cdot j} + \beta_{2,a} \overline{M}_{\cdot(-j)} + \boldsymbol{\beta}_{3,a}^T \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{5,a} N} \Big| A = a^*, \boldsymbol{C}, N \right] \\
=& e^{\beta_{0,a} + \boldsymbol{\beta}_{3,a}^T \boldsymbol{V} + \boldsymbol{\beta}_{4,a}^T \boldsymbol{X}_{\cdot j} + \beta_{5,a} N} \mathbb{E}\left[e^{\beta_{1,a} M_{\cdot j} + \beta_{2,a} \overline{M}_{\cdot(-j)}} \Big| A = a^*, \boldsymbol{C}, N \right] \\
=& \exp\Big\{ \beta_{0,a} + (\beta_{1,a} + \beta_{2,a})\gamma_0 + (\beta_{1,a} + \beta_{2,a})\gamma_1 a^* + 0.5\sigma_M^2 \{\beta_{1,a}^2 + (2\beta_{1,a}\beta_{2,a} + \beta_{2,a}^2)\rho\} \\
& + (\boldsymbol{\beta}_{3,a}^T + (\beta_{1,a} + \beta_{2,a})\boldsymbol{\gamma}_3^T)\boldsymbol{V} + (\boldsymbol{\beta}_{4,a}^T + \beta_{1,a}\boldsymbol{\gamma}_4^T)\boldsymbol{X}_{\cdot j} + \beta_{2,a}\boldsymbol{\gamma}_4^T \overline{\boldsymbol{X}}_{\cdot(-j)} \\
& + ((\beta_{1,a} + \beta_{2,a})(\gamma_2 a^* + \gamma_5) + \beta_{5,a})N + 0.5\sigma_M^2 \beta_{2,a}^2 (1-\rho) \frac{1}{N-1} \Big\}
\end{aligned}$$

This suggests that $\log\left\{\eta_{\cdot j}^{\star}(a, a^*, \boldsymbol{C}, N)\right\} = \zeta_{0,(a,a^*)} + \boldsymbol{\zeta}_{1,(a,a^*)}^T \boldsymbol{V} + \boldsymbol{\zeta}_{2,(a,a^*)}^T \boldsymbol{X}_{\cdot j} + \boldsymbol{\zeta}_{3,(a,a^*)}^T \overline{\boldsymbol{X}}_{\cdot(-j)} + \zeta_{4,(a,a^*)} N + \zeta_{5,(a,a^*)} \frac{1}{N-1}$, where the coefficients can be estimated by running a quasi-Poisson regression based on a log-link function with $\widehat{\eta}_{ij}(a, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ as the response and $\{\boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}, N_i, \frac{1}{N_i-1}\}$ as predictors, based on all clusters with treatment $A = a^*$.

2^*. Parametric models for $\eta^{\dagger}_{\cdot j}(a, a^*, m_{\cdot j}, \boldsymbol{c}, n)$ with a binary/count outcome

By the definition of $\eta^{\dagger}_{\cdot j}$, we have that

$$\begin{aligned}\eta^{\dagger}_{\cdot j}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N) =& \int \eta_{\cdot j}(a, M_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N) f(\boldsymbol{m}_{\cdot(-j)}|A = a^*, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}_{\cdot(-j)} \\ =& e^{\beta_{0,a}+\beta_{1,a}M_{\cdot j}+\boldsymbol{\beta}_{3,a}^T\boldsymbol{V}+\boldsymbol{\beta}_{4,a}^T\boldsymbol{X}_{\cdot j}+\beta_{5,a}N}\mathbb{E}[e^{\frac{\beta_{2,a}}{N-1}\sum_{l\neq j}M_{\cdot l}}|A = a^*, \boldsymbol{C}, N],\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}[e^{\frac{\beta_{2,a}}{N-1}\sum_{l\neq j}M_{\cdot l}}|A = a^*, \boldsymbol{C}, N] \\ =& \exp\left\{\beta_{2,a}\left(\gamma_0 + \gamma_1 a^* + \gamma_2 a^* N + \boldsymbol{\gamma}_3^T\boldsymbol{V} + \boldsymbol{\gamma}_4^T\overline{\boldsymbol{X}}_{\cdot(-j)} + \gamma_5 N\right) + 0.5\sigma_M^2\frac{\beta_{2,a}^2}{(N-1)}(1+(N-2)\rho)\right\}\end{aligned}$$

Therefore, we have

$$\begin{aligned}\eta^{\dagger}_{\cdot j}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N) =& \exp\Big\{\beta_{0,a} + \beta_{2,a}(\gamma_0+\gamma_1 a^*) + 0.5\sigma_M^2\beta_{2,a}^2\rho + \beta_{1,a}M_{\cdot j} + \left(\beta_{2,a}\boldsymbol{\gamma}_3^T + \boldsymbol{\beta}_{3,a}^T\right)\boldsymbol{V} \\ &+ \boldsymbol{\beta}_{4,a}^T\boldsymbol{X}_{\cdot j} + \beta_{2,a}\boldsymbol{\gamma}_4^T\overline{\boldsymbol{X}}_{\cdot(-j)} + (\beta_{2,a}\gamma_5 + \beta_{2,a}\gamma_2 a^* + \beta_{5,a})N + \frac{0.5\sigma_M^2\beta_{2,a}^2\rho}{(N-1)}\Big\}.\end{aligned}$$

This suggests that $\log\left\{\eta^{\dagger}_{\cdot j}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N)\right\} = \psi_{0,(a,a^*)} + \psi_{1,(a,a^*)}M_{\cdot j} + \boldsymbol{\psi}_{2,(a,a^*)}^T\boldsymbol{V} + \boldsymbol{\psi}_{3,(a,a^*)}^T\boldsymbol{X}_{\cdot j} + \boldsymbol{\psi}_{4,(a,a^*)}^T\overline{\boldsymbol{X}}_{\cdot(-j)} + \psi_{5,(a,a^*)}N + \psi_{6,(a,a^*)}\frac{1}{N-1}$, where unknown coefficients can be estimated by running quasi-Poisson regression based on a log-link function with $\frac{\widehat{\eta}_{ij}(a,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)\widehat{\kappa}_{ij}(a,M_{ij},\boldsymbol{C}_i,N_i)}{\widehat{\kappa}^{\star}_{ij}(a^*,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i)}$ as response and $\{M_{ij}, \boldsymbol{V}_i, \boldsymbol{X}_{ij}, \overline{\boldsymbol{X}}_{i(-j)}, N_i, \frac{1}{N_i-1}\}$ as predictors, based on all clusters with treatment $A = a^*$.

□

A.4 Estimating the causal mediation effect estimands

We first derive the EIF of the causal mediation effect estimands. Notice that all of the mediation effects can be defined as $\Delta = g(\zeta_1, \zeta_2)$, with $\zeta_1 \neq \zeta_2 \in \{\theta_C(1,1), \theta_C(1,0), \theta_C(0,0), \theta_I(1,1), \theta_I(1,0), \theta_I(0,0), \tau_C, \tau_I\}$ and $g(\cdot,\cdot)$ as a user-specified function determining the scale of effect measure. Denote $\mathcal{D}_{\zeta_1}(\boldsymbol{O})$ and $\mathcal{D}_{\zeta_2}(\boldsymbol{O})$ as the EIFs of ζ_1 and ζ_2, respectively, where the explicit expressions are provided in Theorem 2. The following Corollary presents the EIF for the general causal estimand Δ. We define $\dot{g}_1(x, y)$ and $\dot{g}_2(x, y)$ as the partial derivative of $g(x, y)$ on x and y, respectively; for example, if we define mediation effects based on a mean difference scale with $g(x, y) = x - y$, then $\dot{g}_1(x, y) = 1$ and $\dot{g}_1(x, y) = -1$.

Corollary 1. *Suppose all conditions in Theorem 2 hold. The EIF of $\Delta = g(\zeta_1, \zeta_2)$ is*

$$\mathcal{D}_{\Delta}(\boldsymbol{O}) = \dot{g}_1(\zeta_1,\zeta_2)\mathcal{D}_{\zeta_1}(\boldsymbol{O}) + \dot{g}_2(\zeta_1,\zeta_2)\mathcal{D}_{\zeta_2}(\boldsymbol{O}).$$

Therefore, the semiparametric efficiency lower bound for estimating Δ is $\mathbb{E}[\{\mathcal{D}_{\Delta}(\boldsymbol{O})\}^2]$.

For any approach $d \in \{\text{mf-par}, \text{eif}_1\text{-par}, \text{eif}_2\text{-par}, \text{eif}_2\text{-ml}, \text{eif}_2\text{-ml}\}$, plug-in estimates of the NIE_V, NDE_V, SME_V, and IME_V can be obtained by $\widehat{\text{NIE}}_V^d = g\left(\widehat{\theta}_V^d(1,1), \widehat{\theta}_V^d(1,0)\right)$,

$\widehat{\mathrm{NDE}}_V^d = \left(\widehat{\theta}_V^d(1,0), \widehat{\theta}_V^d(0,0)\right)$, $\widehat{\mathrm{SME}}_V^d = g\left(\widehat{\theta}_V^d(1,1), \widehat{\tau}_V^d\right)$, and $\widehat{\mathrm{IME}}_V^d = g\left(\widehat{\tau}_V^d, \widehat{\theta}_V^d(1,0)\right)$. We summarize properties of EIF-based estimators of the mediation effects when nuisance functions are estimated based on parametric working models.

Corollary 2. *Suppose all conditions in Theorem 3 hold. The following asymptotic properties hold for both $V \in \{I, C\}$.*

(i) $\widehat{NIE}_V^{\textit{eif}_1\textit{-par}}$ *and* $\widehat{NDE}_V^{\textit{eif}_1\textit{-par}}$ *is CAN if either $\{\kappa_{\cdot j}, \mathcal{C}\}$ or $\eta_{\cdot j}$ are correctly specified, which are semiparametrically efficient when $\{\eta_{\cdot j}, \kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified.* $\widehat{SME}_V^{\textit{eif}_1\textit{-par}}$ *and* $\widehat{IME}_V^{\textit{eif}_1\textit{-par}}$ *are CAN if either $\{\kappa_{\cdot j}, \mathcal{C}\}$ or $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ are correctly specified, which are semiparametrically efficient when $\{\eta_{\cdot j}, \kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified.*

(ii) $\widehat{NIE}_V^{\textit{eif}_2\textit{-par}}$ *and* $\widehat{NDE}_V^{\textit{eif}_2\textit{-par}}$ *are CAN if either $\eta_{\cdot j}$ or s is correctly specified, which are semiparametrically efficient when $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, s\}$ are correctly specified.* $\widehat{SME}_V^{\textit{eif}_2\textit{-par}}$ *and* $\widehat{IME}_V^{\textit{eif}_2\textit{-par}}$ *are CAN if either $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ or $\{\kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ are correctly specified, which are semiparametrically efficient when $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ are correctly specified.*

The form of Corollary 2 shows that robustness property of NIE and NDE estimates are identical to the robustness property for estimates of $\theta_V(a, a^*)$ in Theorem 3, and robustness property of SME and IME estimates are identical to the robustness property of estimates of τ_V. The cluster bootstrap can be borrowed for purpose of inference for all estimators considered in Corollary 2. Corollary 3 deduces the properties of the EIF-based estimates of mediation effects when nuisance is estimated by machine learners.

Corollary 3. *Suppose all conditions in Theorem 4 hold. The following asymptotic properties hold for both $V \in \{I, C\}$.*

(i) $\widehat{NIE}_V^{\textit{eif}_1\textit{-ml}}$ *and* $\widehat{NDE}_V^{\textit{eif}_1\textit{-ml}}$ *are consistent if $\widehat{\kappa}^c$ or $\widehat{\eta}_{\cdot j}$ is consistent in $L_2(P)$-norm, which are CAN and semiparametrically efficient if $\{\widehat{\kappa}^c, \widehat{\eta}_{\cdot j}\}$ are consistent with $\lambda_K(\widehat{\kappa}^c)\lambda_K(\widehat{\eta}_{\cdot j}) = o(K^{-1/2})$.* $\widehat{SME}_V^{\textit{eif}_1\textit{-ml}}$ *and* $\widehat{IME}_V^{\textit{eif}_1\textit{-ml}}$ *are consistent if $\{\widehat{\kappa}_{\cdot j}, \widehat{\eta}_{\cdot j}\}$ or $\{\widehat{\kappa}_{\cdot j}, \widehat{\kappa}^c_{\cdot(-j)}, \widehat{\kappa}^c\}$ is consistent in $L_2(P)$-norm, which are CAN and semiparametrically efficient if $\{\widehat{\kappa}_{\cdot j}, \widehat{\eta}_{\cdot j}, \widehat{\kappa}^c_{\cdot(-j)}, \widehat{\kappa}^c\}$ are consistent with $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^c_{\cdot(-j)}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^c) + \lambda_K(\widehat{\kappa}^c_{\cdot(-j)})\lambda_K(\widehat{\kappa}_{\cdot j}) = o(K^{-1/2})$.*

(ii) $\widehat{NIE}_V^{\textit{eif}_2\textit{-ml}}$ *and* $\widehat{NDE}_V^{\textit{eif}_2\textit{-ml}}$ *are consistent if $\widehat{s}$ or $\widehat{\eta}_{\cdot j}$ is consistent in $L_2(P)$-norm, which are CAN and semiparametrically efficient if $\{\widehat{\eta}_{\cdot j}, \widehat{\eta}^\star_{\cdot j}, \widehat{s}\}$ are consistent with $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{s}) = o(K^{-1/2})$.* $\widehat{SME}_V^{\textit{eif}_2\textit{-ml}}$ *and* $\widehat{IME}_V^{\textit{eif}_2\textit{-ml}}$ *are consistent if $\{\widehat{\kappa}_{\cdot j}, \widehat{\eta}_{\cdot j}\}$ or $\{\widehat{\kappa}_{\cdot j}, \widehat{\kappa}^\star_{\cdot j}, \widehat{s}\}$ are consistent in $L_2(P)$-norm, and are CAN and semiparametrically efficient if $\{\widehat{\eta}_{\cdot j}, \widehat{\eta}^\star_{\cdot j}, \widehat{\eta}^\dagger_{\cdot j}, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^\star_{\cdot j}, \widehat{s}\}$ are consistent with $\lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{\kappa}^\star_{\cdot j}) + \lambda_K(\widehat{\eta}_{\cdot j})\lambda_K(\widehat{s}) + \lambda_K(\widehat{\eta}^\dagger_{\cdot j})\lambda_K(\widehat{\kappa}_{\cdot j}) = o(K^{-1/2})$.*

For purpose of inference for the machine learning estimators, we can approximate asymptotic variances by the empirical variance of their corresponding EIFs, where the expressions of the EIF are shown in Corollary 1. For example, the variance of $\widehat{\mathrm{NIE}}_C^{\mathrm{eif}_1\text{-ml}}$ can be estimated by

$$\mathrm{Var}\left(\widehat{\mathrm{NIE}}_C^{\mathrm{eif}_1\text{-ml}}\right) = \frac{1}{K}\mathbb{P}_K\Big[\Big\{\dot{g}_1(\widehat{\theta}_C^{\mathrm{eif}_1\text{-ml}}(1,1), \widehat{\theta}_C^{\mathrm{eif}_1\text{-ml}}(1,0)) \times \widehat{\mathcal{D}}_{\theta_C(1,1)}(\boldsymbol{O})$$

$$+ \dot{g}_2(\widehat{\theta}_C^{\text{eif}_1\text{-ml}}(1,1), \widehat{\theta}_C^{\text{eif}_1\text{-ml}}(1,0)) \times \widehat{\mathcal{D}}_{\theta_C(1,0)}(\boldsymbol{O})\Big\}^2\Big],$$

where $\widehat{\mathcal{D}}_{\theta_C(a,a^*)}(\boldsymbol{O}) = \widehat{\psi}_\theta(a, a^*; \boldsymbol{O}_i) - \widehat{\theta}_C^{\text{eif}_1\text{-ml}}(a, a^*)$ is the EIF of $\theta_C(a, a^*)$ evaluated based on $\widehat{h}^{(1)}_{nuisance}$.

A.5 Supporting information on the simulation and application studies

We first describe the data generation process in the simulation study. We simulate 1000 cluster-randomized experiments, each with $K = 100$ clusters. Data in each cluster is generated according to the following process. First, the cluster size N_i is drawn from a discrete uniform distribution on $\{10, \ldots, 50\}$. We then generate a cluster-level covariate $V_i \sim N(\frac{3N_i}{50}, 1)$, an individual-level covariate $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{iN_i}]^T$ with each X_{ij} independently drawn from $N(2V_i, 1)$. Next, we randomize the treatment $A_i \sim \text{Bernoulli}(0.5)$. The mediator $\boldsymbol{M}_i = [M_{i1}, \ldots, M_{iN_i}]^T$ is simulated based on (s1), where marginally $M_{ij}|\{A_i, \boldsymbol{X}_i, V_i, N_i\} \sim N(-2 + 0.2A_i + \frac{(0.5+0.5A_i)N_i}{50} + 0.5X_{ij} + 0.5V_i, 2^2)$ and $\mathcal{C}$ is a Gaussian copula based on an exchangeable association structure with $\rho = 0.1$. We simulate the outcome $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iN_i}]^T$ based on the joint distribution

$$P_{\boldsymbol{Y}_i|A_i,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i}(\boldsymbol{y}_i|A_i, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i) = \mathcal{C}(F_{i1}(y_{i1}), \ldots, F_{iN_i}(y_{iN_i})|A_i, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i),$$

where $F_{ij}(y_{ij}) \equiv P_{Y_{ij}|A_i,\boldsymbol{M}_i,\boldsymbol{C}_i,N_i}(y_{ij}|A_i, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i)$ is the marginal distribution of Y_{ij} and $\mathcal{C}$ is a copula. We set $\mathcal{C}$ as a Gaussian copula based on an exchangeable association structure with $\rho = 0.1$ and marginally $Y_{ij}|\{A_i, \boldsymbol{M}_i, \boldsymbol{C}_i, N_i\} \sim N(A_i + \frac{(0.5+0.5A_i)N_i}{50} + \frac{0.5}{N_i-1}\sum_{l\neq j} M_{il} + M_{ij} + 0.5X_{ij} + 0.5V_i, 2^2)$.

For both of the simulation and application studies, we consider the following parametric working models in $h^{(2)}_{nuisance}$:

1. $\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)$: same to (s5).
2. $\eta^\star_{\cdot j}(a, a^*, \boldsymbol{C}, N) = \zeta_{0,(a,a^*)} + \boldsymbol{\zeta}^T_{1,(a,a^*)}\boldsymbol{V} + \boldsymbol{\zeta}^T_{2,(a,a^*)}\boldsymbol{X}_{\cdot j} + \boldsymbol{\zeta}^T_{3,(a,a^*)}\overline{\boldsymbol{X}}_{\cdot(-j)} + \zeta_{4,(a,a^*)}N$
3. $\eta^\dagger_{\cdot j}(a, a^*, M_{\cdot j}, \boldsymbol{C}, N) = \psi_{0,(a,a^*)} + \psi_{1,(a,a^*)}M_{\cdot j} + \boldsymbol{\psi}^T_{2,(a,a^*)}\boldsymbol{V} + \boldsymbol{\psi}^T_{3,(a,a^*)}\boldsymbol{X}_{\cdot j} + \boldsymbol{\psi}^T_{4,(a,a^*)}\overline{\boldsymbol{X}}_{\cdot(-j)} + \psi_{5,(a,a^*)}N$
4. $\kappa_{\cdot j}(A, M_{\cdot j}, \boldsymbol{C}, N)$: same to (s2)
5. $\kappa^\star_{\cdot j}(A, \boldsymbol{M}, \boldsymbol{C}, N)$ follows a normal distribution with mean (s6) and variance $\sigma^{\star 2}_M$.
6. $s(A, \boldsymbol{M}, \boldsymbol{C}, N)$ follows the logistic regression (s7).

For misspecified models, we remove all baseline covariates from the working regressions (e.g., we remove $\{\boldsymbol{V}, \boldsymbol{X}_{\cdot j}\}$ in the working model for $\eta_{\cdot j}$ under misspecification). In $h^{(1)}_{nusiance}$, one also needs to model the copula $\mathcal{C}$, which is considered as a Gaussian copula with exchangeable association structures; under misspecification, we further fix $\rho = 0$.

B Proofs of the Theorems, Propositions, Remarks, and Corollaries

Proof of Theorem 1. We can show

$$\theta_C(a,a^*) = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(a, M(a^*)\right)\right]$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \mathbb{E}\left\{Y_{\cdot j}\left(a, M(a^*)\right) \Big| \boldsymbol{C}, N\right\}\right] \quad \text{(by law of iterated expectations (LIE))}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a, \boldsymbol{m}\right) | \boldsymbol{M}(a^*) = \boldsymbol{m}, \boldsymbol{C}, N\right\} \mathbb{P}(\boldsymbol{M}(a^*) = \boldsymbol{m} | \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\right] \quad \text{(by LIE)}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a, \boldsymbol{m}\right) | A = a, \boldsymbol{M}(a^*) = \boldsymbol{m}, \boldsymbol{C}, N\right\} \mathbb{P}(\boldsymbol{M}(a^*) = \boldsymbol{m} | A = a^*, \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\right]$$

(by Assumption 2)

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a, \boldsymbol{m}\right) | A = a, \boldsymbol{M}(a) = \boldsymbol{m}, \boldsymbol{C}, N\right\} \mathbb{P}(\boldsymbol{M}(a^*) = \boldsymbol{m} | A = a^*, \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\right]$$

(by Assumption 3)

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j} | A = a, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{C}, N\right\} \mathbb{P}(\boldsymbol{M} = \boldsymbol{m} | A = a^*, \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\right]$$

(by Assumption 1)

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N) \kappa(a^*, \boldsymbol{m}, \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\right].$$

and

$$\tau_C = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(1, M_{\cdot j}(1), \boldsymbol{M}_{\cdot(-j)}(0)\right)\right]$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \mathbb{E}\left\{Y_{\cdot j}\left(1, M_{\cdot j}(1), \boldsymbol{M}_{\cdot(-j)}(0)\right) \Big| \boldsymbol{C}, N\right\}\right] \quad \text{(by LIE)}$$

$$=\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right) | M_{\cdot j}(1) = m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}(0) = \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N\right\}$$

$$\mathbb{P}(M_{\cdot j}(1) = m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}(0) = \boldsymbol{m}_{\cdot(-j)} | \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by LIE)}$$

$$=\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N} \int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right) | M_{\cdot j}(1) = m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}(0) = \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N\right\}$$

$$\mathbb{P}(M_{\cdot j}(1) = m_{\cdot j} | \boldsymbol{C}, N) \mathbb{P}(\boldsymbol{M}_{\cdot(-j)}(0) = \boldsymbol{m}_{\cdot(-j)} | \boldsymbol{C}, N) \mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by Assumption 5)}$$

$$=\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right)|A=1, M_{\cdot j}(1)=m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N\right\}$$

$$\mathbb{P}(M_{\cdot j}(1)=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by Assumption 2)}$$

$$=\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right)|A=1, M_{\cdot j}(1)=m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}(1)=\boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N\right\}$$

$$\mathbb{P}(M_{\cdot j}(1)=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by Assumption 3)}$$

$$=\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right)|A=1, M_{\cdot j}=m_{\cdot j}, \boldsymbol{M}_{\cdot(-j)}=\boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N\right\}$$

$$\mathbb{P}(M_{\cdot j}=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by Assumption 1)}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right].$$

Following a similar procedure, one can derive the identification formulas for $\theta_I(a, a^*)$ and τ_I. □

Supporting information for Remark 1. We can show that the identification formulas in Theorem 1 have causal interpretations under the interventional causal mediation framework, under Assumptions 1–4. Specifically, let $\boldsymbol{G}(1)$ be a random draw of $\boldsymbol{M}(1)$ based on the true distribution of $\boldsymbol{M}(1)$ conditional on baseline information, i.e., $P_{\boldsymbol{M}(1)|\boldsymbol{C},N}$. Also let $\boldsymbol{G}(0)$ be a random draw of $\boldsymbol{M}(0)$ based on $P_{\boldsymbol{M}(0)|\boldsymbol{C},N}$. Notice that $\boldsymbol{G}(1) \perp \boldsymbol{G}(0)|\boldsymbol{C}, N$ due to our construction, which further implies

$$G_{\cdot j}(1) \perp \boldsymbol{G}_{\cdot(-j)}(0)|\boldsymbol{C}, N, \text{ for any } j. \tag{s9}$$

Also by our construction of $\boldsymbol{G}(a)$, we have that

$$\{\boldsymbol{G}(1), \boldsymbol{G}(0)\} \perp Y_{\cdot j}(a, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)})|\{A, \boldsymbol{C}, N\} \tag{s10}$$

for any j, $a \in \{0, 1\}$ and $\{m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\}$ over their valid support. Define the following interventional counterparts for $\theta_C(a, a^*)$, $\theta_V(a, a^*)$, τ_C, and τ_V, respectively:

$$\theta_C^{(\text{int})}(a, a^*) = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(a, G_{\cdot j}(a^*), \boldsymbol{G}_{\cdot(-j)}(a^*)\right)\right],$$

$$\theta_I^{(\text{int})}(a, a^*) = \mathbb{E}\left[\sum_{j=1}^{N} Y_{\cdot j}\left(a, G_{\cdot j}(a^*), \boldsymbol{G}_{\cdot(-j)}(a^*)\right)\right]/\mathbb{E}[N],$$

$$\tau_C^{(\text{int})} = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(1, G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(0)\right)\right],$$

$$\tau_I^{(\text{int})} = \mathbb{E}\left[\sum_{j=1}^{N} Y_{\cdot j}\left(1, G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(0)\right)\right] / \mathbb{E}[N].$$

Next, we show that the mediation formulas in Theorem 1 exactly coincide with $\theta_C^{(\text{int})}(a,a^*)$, $\theta_I^{(\text{int})}(a,a^*)$, $\tau_C^{(\text{int})}$, and $\tau_I^{(\text{int})}$, respectively. Specifically,

$$\begin{aligned}
\theta_C^{(\text{int})}(a,a^*) &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(a, \boldsymbol{G}(a^*)\right)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \mathbb{E}\left\{Y_{\cdot j}\left(a, \boldsymbol{G}(a^*)\right) \Big| \boldsymbol{C}, N\right\}\right] \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a,\boldsymbol{m}\right)|\boldsymbol{G}(a^*)=\boldsymbol{m},\boldsymbol{C},N\right\}\mathbb{P}(\boldsymbol{G}(a^*)=\boldsymbol{m}|\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \quad \text{(by LIE)} \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a,\boldsymbol{m}\right)|\boldsymbol{G}(a^*)=\boldsymbol{m},\boldsymbol{C},N\right\}\mathbb{P}(\boldsymbol{G}(a^*)=\boldsymbol{m}|\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a,\boldsymbol{m}\right)|A=a,\boldsymbol{G}(a^*)=\boldsymbol{m},\boldsymbol{C},N\right\}\mathbb{P}(\boldsymbol{M}(a^*)=\boldsymbol{m}|A=a^*,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&\text{(by Assumption 2 and the construction of } \boldsymbol{G}(a^*)) \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(a,\boldsymbol{m}\right)|A=a,\boldsymbol{C},N\right\}\mathbb{P}(\boldsymbol{M}(a^*)=\boldsymbol{m}|A=a^*,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&\text{(by equation (s10))} \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}|A=a,\boldsymbol{M}=\boldsymbol{m},\boldsymbol{C},N\right\}\mathbb{P}(\boldsymbol{M}=\boldsymbol{m}|A=a^*,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&\text{(by Assumptions 1 and 3)} \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\kappa(a^*,\boldsymbol{m},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
=&\text{the identification formula of } \theta_C(a,a^*) \text{ in Theorem 1.}
\end{aligned}$$

and

$$\begin{aligned}
\tau_C^{(\text{int})} &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} Y_{\cdot j}\left(1, G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(0)\right)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} \mathbb{E}\left\{Y_{\cdot j}\left(1, G_{\cdot j}(1), \boldsymbol{G}_{\cdot(-j)}(0)\right) \Big| \boldsymbol{C}, N\right\}\right] \\
=&\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right)|G_{\cdot j}(1)=m_{\cdot j},\boldsymbol{G}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N\right\} \\
&\quad \mathbb{P}(G_{\cdot j}(1)=m_{\cdot j},\boldsymbol{G}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by LIE)} \\
=&\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}\left\{Y_{\cdot j}\left(1, m_{\cdot j}, \boldsymbol{m}_{\cdot(-j)}\right)|G_{\cdot j}(1)=m_{\cdot j},\boldsymbol{G}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N\right\} \\
&\quad \mathbb{P}(G_{\cdot j}(1)=m_{\cdot j}|\boldsymbol{C},N)\mathbb{P}(\boldsymbol{G}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\Bigg] \quad \text{(by equation (s9))}
\end{aligned}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1,m_{\cdot j},\boldsymbol{m}_{\cdot(-j)}\right)|A=1,\boldsymbol{C},N\right\}\right.$$

$$\left.\mathbb{P}(M_{\cdot j}(1)=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \quad \text{(by Assumption 2 and (s10))}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1,m_{\cdot j},\boldsymbol{m}_{\cdot(-j)}\right)|A=1,M_{\cdot j}(1)=m_{\cdot j},\boldsymbol{M}_{\cdot(-j)}(1)=\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N\right\}\right.$$

$$\left.\mathbb{P}(M_{\cdot j}(1)=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}(0)=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \quad \text{(by Assumption 3)}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\mathbb{E}\left\{Y_{\cdot j}\left(1,m_{\cdot j},\boldsymbol{m}_{\cdot(-j)}\right)|A=1,M_{\cdot j}=m_{\cdot j},\boldsymbol{M}_{\cdot(-j)}=\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N\right\}\right.$$

$$\left.\mathbb{P}(M_{\cdot j}=m_{\cdot j}|A=1,\boldsymbol{C},N)\mathbb{P}(\boldsymbol{M}_{\cdot(-j)}=\boldsymbol{m}_{\cdot(-j)}|A=0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \quad \text{(by Assumption 1)}$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]$$

=the identification formula of τ_C in Theorem 1.

Following a similar procedure, one can show the identification formulas of $\theta_I(a,a^*)$ and τ_I align with $\theta_I^{(\text{int})}(a,a^*)$ and $\tau_I^{(\text{int})}$. □

Proof of Theorem 2. To simplify the notations, we shall suppress the subscript in the probability density functions. Let $f(\boldsymbol{O})\equiv f(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)$ be the joint density function of $\boldsymbol{O}$, where we suppress the subscript in the density function f for ease of notations. The joint density function of observed data can be decomposed as

$$\begin{aligned}f(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)&=f(\boldsymbol{Y}|A,\boldsymbol{M},\boldsymbol{C},N)f(\boldsymbol{M}|A,\boldsymbol{C},N)f(A|\boldsymbol{C},N)f(\boldsymbol{C},N)\\&=f(\boldsymbol{Y}|A,\boldsymbol{M},\boldsymbol{C},N)f(\boldsymbol{M}|A,\boldsymbol{C},N)f(A)f(\boldsymbol{C},N),\end{aligned}$$

where $f(A|\boldsymbol{C},N)=f(A)$ due to randomization (Assumption 2). Except that $f(A)$ is known due to study design, all other components in $f(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)$ are unknown. Consider the following parametric submodel for $f(\boldsymbol{O})$ index by a one-dimensional parameter $t\in(-\epsilon,\epsilon)$ with certain $\epsilon>0$:

$$f_t(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)=f_t(\boldsymbol{Y}|A,\boldsymbol{M},\boldsymbol{C})f_t(\boldsymbol{M}|A,\boldsymbol{C})f(A)f_t(\boldsymbol{C},N),$$

where $f_0(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)=f(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)$.

We first derive the EIF of $\theta_C(a,a^*)$. The value of $\theta_C(a,a^*)$ evaluated under the parametric submodel is

$$\theta_C(a,a^*;t)=\int_n\frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{c},\boldsymbol{m}}\mathbb{E}_t[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]f_t(\boldsymbol{m}|a^*,\boldsymbol{c})f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n,$$

where $\mathbb{E}_t$ is the expectation operator under the parametric submodel. Notice that $\theta_C(a,a^*;0)\equiv$

$\theta_C(a,a^*)$. The EIF of $\theta_C(a,a^*)$ is the unique function $\mathcal{D}_{\theta_C(a,a^*)}(\boldsymbol{O})$ satisfies

$$\nabla_{t=0}\theta_C(a,a^*;t) = \mathbb{E}[\mathcal{D}_{\theta_C(a,a^*)}(\boldsymbol{O})S(\boldsymbol{O})],$$

where $\nabla_{t=0}$ is the pathwise derivative with respect to t at $t=0$, and $S(\boldsymbol{O})$ is the score function of the parametric submodel. Taking the pathwise derivative of $\theta_C(a,a^*;t)$ with respect to t leads to

$$\begin{aligned}
&\nabla_{t=0}\theta_C(a,a^*;t)\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m}} \nabla_{t=0}\mathbb{E}_t[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n] f(\boldsymbol{m}|a^*,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n && \text{(s11)}\\
&\quad + \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]\nabla_{t=0} f_t(\boldsymbol{m}|a^*,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n && \text{(s12)}\\
&\quad + \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n] f(\boldsymbol{m}|a^*,\boldsymbol{c},n)\nabla_{t=0} f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n && \text{(s13)}
\end{aligned}$$

where

$$\begin{aligned}
\text{(s11)} &= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m}} \nabla_{t=0}\mathbb{E}_t[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)} f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} y_{\cdot j}\nabla_{t=0} f_t(\boldsymbol{y}|a,\boldsymbol{m},\boldsymbol{c},n)\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)} f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} y_{\cdot j}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)}\{S(\boldsymbol{y},\boldsymbol{m},a,\boldsymbol{c},n) - \mathbb{E}[S(\boldsymbol{Y},\boldsymbol{m},a,\boldsymbol{c},n)|\boldsymbol{m},a,\boldsymbol{c},n]\}\\
&\qquad f(\boldsymbol{y}|a,\boldsymbol{m},\boldsymbol{c},n) f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} y_{\cdot j}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)} S(\boldsymbol{y},\boldsymbol{m},a,\boldsymbol{c},n) f(\boldsymbol{y}|a,\boldsymbol{m},\boldsymbol{c},n) f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&\quad - \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} y_{\cdot j}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)}\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{m},a,\boldsymbol{c},n)|\boldsymbol{m},a,\boldsymbol{c},n] f(\boldsymbol{y}|a,\boldsymbol{m},\boldsymbol{c},n) f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} \{y_{\cdot j} - \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]\}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)} S(\boldsymbol{y},\boldsymbol{m},a,\boldsymbol{c},n) f(\boldsymbol{y}|a,\boldsymbol{m},\boldsymbol{c},n) f(\boldsymbol{m}|a,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \mathbb{E}\left[\left\{\frac{1}{N}\sum_{j=1}^N \frac{\mathbb{I}(A=a)}{f(A=a)}\frac{f(\boldsymbol{M}|a^*,\boldsymbol{C},N)}{f(\boldsymbol{M}|a,\boldsymbol{C},N)}\{Y_{\cdot j} - \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{M},\boldsymbol{C},N]\}\right\} S(\boldsymbol{O})\right]
\end{aligned}$$

and

$$\begin{aligned}
\text{(s12)} &= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]\{\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{m},a^*,\boldsymbol{c},n)|\boldsymbol{m},a^*,\boldsymbol{c},n] - \mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},a^*,\boldsymbol{c},n)|a^*,\boldsymbol{c},n]\}\\
&\qquad f(\boldsymbol{m}|a^*,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&= \int_n \frac{1}{n}\sum_{j=1}^n \int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n] S(\boldsymbol{y},\boldsymbol{m},a^*,\boldsymbol{c},n) f(\boldsymbol{y}|a^*,\boldsymbol{m},\boldsymbol{c},n) f(\boldsymbol{m}|a^*,\boldsymbol{c},n) f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n
\end{aligned}$$

$$- \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}}\int_{\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]f(\boldsymbol{m}|a^*,\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\times S(\boldsymbol{y},\boldsymbol{m},a^*,\boldsymbol{c},n)f(\boldsymbol{y}|a^*,\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m}|a^*,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n$$

$$= \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{c},\boldsymbol{m},\boldsymbol{y}} \{\mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n] - u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{c},n)\}S(\boldsymbol{y},\boldsymbol{m},a^*,\boldsymbol{c},n)f(\boldsymbol{y}|a^*,\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m}|a^*,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n$$

$$=\mathbb{E}\left[\left\{\frac{1}{N}\sum_{j=1}^{N}\{\mathbb{E}[Y_{\cdot j}|a,\boldsymbol{M},\boldsymbol{C},N] - u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N)\}\frac{\mathbb{I}(A=a^*)}{f(A=a^*)}\right\}S(\boldsymbol{O})\right]$$

and

$$(\text{s13}) = \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{c},\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n]f(\boldsymbol{m}|a^*,\boldsymbol{c},n)\{\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{c},n)|\boldsymbol{c},n] - \mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)]\}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{C},N]f(\boldsymbol{m}|a^*,\boldsymbol{C},N)d\boldsymbol{m}S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right] - \theta_C(a,a^*)\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)]$$

$$=\mathbb{E}\left[\left\{\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{C},N]f(\boldsymbol{m}|a^*,\boldsymbol{C},N)d\boldsymbol{m} - \theta_C(a,a^*)\right\}S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right]$$

$$=\mathbb{E}\left[\left\{\frac{1}{N}\sum_{j=1}^{N}u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N) - \theta_C(a,a^*)\right\}S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right]$$

It follows that

$$\begin{aligned}\mathcal{D}_{\theta_C(a,a^*)}(\boldsymbol{O}) =&\frac{1}{N}\sum_{j=1}^{N}\left\{\frac{\mathbb{I}(A=a)}{f(A=a)}\frac{f(\boldsymbol{M}|a^*,\boldsymbol{C},N)}{f(\boldsymbol{M}|a,\boldsymbol{C},N)}\{Y_{\cdot j} - \mathbb{E}[Y_{\cdot j}|a,\boldsymbol{M},\boldsymbol{C},N]\}\right.\\ &\left.+\frac{\mathbb{I}(A=a^*)}{f(A=a^*)}\{\mathbb{E}[Y_{\cdot j}|a,\boldsymbol{M},\boldsymbol{C},N] - u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N)\} + u_{\cdot j}^{(1)}(a,a^*,\boldsymbol{C},N)\right\} - \theta_C(a,a^*)\\ =&\psi_\theta(a,a^*;\boldsymbol{O}) - \theta_C(a,a^*).\end{aligned}$$

Then, we derive the EIF of τ_C, which is the unique function $\mathcal{D}_{\tau_C}(\boldsymbol{O})$ that satisfies the following equation:

$$\triangledown_{t=0}\tau_C(t) = \mathbb{E}[\mathcal{D}_{\tau_C}(\boldsymbol{O})S_{t=0}(\boldsymbol{O})].$$

Here, $\tau_C(t) = \int_n \frac{1}{n}\sum_{j=1}^{n}\iint_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}_t[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f_t(m_{\cdot j}|1,\boldsymbol{c},n)f_t(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n$. Taking the derivative of $\tau_C(t)$ with respect to t at $t=0$ leads to

$$\triangledown_{t=0}\tau_C(t)$$

$$= \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \triangledown_{t=0}\mathbb{E}_t[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \quad (\text{s14})$$

$$+ \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\triangledown_{t=0}f_t(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \quad (\text{s15})$$

$$+ \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f(m_{\cdot j}|1,\boldsymbol{c},n)\triangledown_{t=0}f_t(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \quad (\text{s16})$$

$$+ \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)\triangledown_{t=0}f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \quad (\text{s17})$$

where

$$
\begin{aligned}
(\text{s14}) =& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \nabla_{t=0}\mathbb{E}_t[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\frac{f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)}{f(\boldsymbol{m}|1,\boldsymbol{c},n)}f(\boldsymbol{m}|1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{y},\boldsymbol{m},\boldsymbol{c}} y_{\cdot j}\nabla_{t=0}f_t(\boldsymbol{y}|1,\boldsymbol{m},\boldsymbol{c},n)\frac{f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)}{f(\boldsymbol{m}|1,\boldsymbol{c},n)}f(\boldsymbol{m}|1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{y},\boldsymbol{m},\boldsymbol{c}} y_{\cdot j}\frac{f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)}{f(\boldsymbol{m}|1,\boldsymbol{c},n)}\left(S(\boldsymbol{y},1,\boldsymbol{m},\boldsymbol{c},n)-\mathbb{E}[S(\boldsymbol{Y},1,\boldsymbol{m},\boldsymbol{c},n)|1,\boldsymbol{m},\boldsymbol{c},n]\right)\\
&\quad\times f(\boldsymbol{y}|1,\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m}|1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_{\boldsymbol{y},\boldsymbol{m},\boldsymbol{c},n} \frac{1}{n}\sum_{j=1}^{n}\left(y_{\cdot j}-\mathbb{E}[y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\right)\frac{f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)}{f(\boldsymbol{m}|1,\boldsymbol{c},n)}S(\boldsymbol{y},1,\boldsymbol{m},\boldsymbol{c},n)\\
&\quad\times f(\boldsymbol{y}|1,\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m}|1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left(Y_{\cdot j}-\mathbb{E}[Y_{\cdot j}|1,\boldsymbol{M},\boldsymbol{C},N]\right)\frac{f(M_j|1,\boldsymbol{C},N)f(\boldsymbol{M}_{\cdot(-j)}|0,\boldsymbol{C},N)}{f(\boldsymbol{M}|1,\boldsymbol{C},N)}\frac{A}{f(A=1)}S(\boldsymbol{Y},A,\boldsymbol{M},\boldsymbol{C},N)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
(\text{s15}) =& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\nabla_{t=0}f_t(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\Big(\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M}_{\cdot(-j)},m_{\cdot j},1,\boldsymbol{c},n)|1,m_{\cdot j},\boldsymbol{c},n]-\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},1,\boldsymbol{c},n)|1,\boldsymbol{c},n]\Big)\\
&\quad\times f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_{\boldsymbol{y},\boldsymbol{m},\boldsymbol{c},n} \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m}_{-j}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}_{-j}S(\boldsymbol{y},\boldsymbol{m},1,\boldsymbol{c},n)f(\boldsymbol{y}|\boldsymbol{m},1,\boldsymbol{c},n)f(m_{\cdot j}|1,\boldsymbol{c},n)\\
&\quad f(\boldsymbol{m}_{\cdot(-j)}|m_{\cdot j},1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
&-\int_{\boldsymbol{y},\boldsymbol{m},\boldsymbol{c},n} \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\times S(\boldsymbol{y},\boldsymbol{m},1,\boldsymbol{c},n)\\
&\quad f(\boldsymbol{y}|\boldsymbol{m},1,\boldsymbol{c},n)f(\boldsymbol{m}|1,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{u_{\cdot j}^{(2)}(1,0,M_j,\boldsymbol{C},N)-u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}\frac{A}{f(A=1)}S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
(\text{s16}) =& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]f_t(m_{\cdot j}|1,\boldsymbol{c},n)\nabla_{t=0}f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=& \int_n \frac{1}{n}\sum_{j=1}^{n}\int_{\boldsymbol{m},\boldsymbol{c}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n]\left(\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{m}_{-j},M_j,0,\boldsymbol{c},n)|0,\boldsymbol{m}_{-j},\boldsymbol{c},n]-\mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},0,\boldsymbol{c},n)|0,\boldsymbol{c},n]\right)\\
&\quad\times f(m_{\cdot j}|1,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n)f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n\\
=&E\left[\frac{1}{N}\sum_{j=1}^{N}\left\{u_{\cdot j}^{(3)}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)-u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}\frac{1-A}{1-f(A=1)}S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
(\text{s17}) = &\int_{\boldsymbol{m},\boldsymbol{c},n} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n] f(m_{\cdot j}|1,\boldsymbol{c},n) f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n) \nabla_{t=0} f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \\
= &\int_{\boldsymbol{m},\boldsymbol{c},n} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{c},n] f(m_{\cdot j}|1,\boldsymbol{c},n) f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{c},n) \left(\mathbb{E}[S(\boldsymbol{Y},A,\boldsymbol{M},\boldsymbol{c},n)|\boldsymbol{c},n] - \mathbb{E}[S(\boldsymbol{Y},A,\boldsymbol{M},\boldsymbol{C},N)]\right) \\
&\times f(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n \\
= &\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \mathbb{E}[Y_{\cdot j}|1,\boldsymbol{m},\boldsymbol{C},N] f(m_{\cdot j}|1,\boldsymbol{C},N) f(\boldsymbol{m}_{\cdot(-j)}|0,\boldsymbol{C},N)\mathrm{d}\boldsymbol{m} S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right] \\
&- \tau_C \mathbb{E}[S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)] \\
= &\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N) - \tau_C\right) S(\boldsymbol{Y},\boldsymbol{M},A,\boldsymbol{C},N)\right].
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathcal{D}_{\tau_C}(\boldsymbol{O}) = &\frac{1}{N}\sum_{j=1}^{N}\Bigg\{\frac{A}{\pi} w^{(2)}(1,0,1,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j} - \eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\} \\
&+ \frac{A}{\pi}\left\{u_{\cdot j}^{(2)}(1,0,M_{\cdot j},\boldsymbol{C},N) - u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\} \\
&+ \frac{1-A}{1-\pi}\left\{u_{\cdot j}^{(3)}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N) - u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\} + u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\Bigg\} - \tau_C
\end{aligned}
$$

Next we show the EIFs of $\theta_I(a,a^*)$ and τ_I. Note that $\theta_I(a,a^*)$ evaluated under the parametric submodel is $\theta_I(a,a^*;t) = \frac{\Psi_I(a,a^*;t)}{\mathbb{E}_t[N]}$, where

$$
\Psi_I(a,a^*;t) = \int_n \sum_{j=1}^{n} \int_{\boldsymbol{c},\boldsymbol{m}} \mathbb{E}_t[Y_{\cdot j}|a,\boldsymbol{m},\boldsymbol{c},n] f_t(\boldsymbol{m}|a^*,\boldsymbol{c}) f_t(\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}\mathrm{d}\boldsymbol{c}\mathrm{d}n.
$$

The EIF of $\theta_I(a,a^*)$ is therefore the unique function $\mathcal{D}_{\theta_I(a,a^*)}(\boldsymbol{O})$ satisfying

$$
\nabla_{t=0}\theta_I(a,a^*;t) = \mathbb{E}[\mathcal{D}_{\theta_I(a,a^*)}(\boldsymbol{O})S(\boldsymbol{O})].
$$

The pathwise derivative of $\theta_I(a,a^*;t)$ in terms of t is

$$
\begin{aligned}
\nabla_{t=0}\theta_I(a,a^*;t) &= \frac{\nabla_{t=0}\Psi_I(a,a^*;t)}{E[N]} - \frac{\Psi_I(a,a^*;0)}{\mathbb{E}[N]^2}\nabla_{t=0}\mathbb{E}_t[N] \\
&= \frac{\nabla_{t=0}\Psi_I(a,a^*;t)}{\mathbb{E}[N]} - \frac{\theta_I(a,a^*)}{\mathbb{E}[N]}\nabla_{t=0}\mathbb{E}_t[N],
\end{aligned}
$$

where it is obvious $\nabla_{t=0}\mathbb{E}_t[N] = \mathbb{E}[(N-\mathbb{E}[N])S(\boldsymbol{O})]$. Moreover, following similar procedure to the derivation of $\nabla_{t=0}\theta_C(a,a^*;t)$, one have

$$
\nabla_{t=0}\Psi_I(a,a^*;t) = E\left[\{N\psi_\theta(a,a^*;\boldsymbol{O}) - \Psi_I(a,a^*;0)\}S(\boldsymbol{O})\right].
$$

This suggests that

$$\begin{aligned}
\nabla_{t=0}\theta_I(a,a^*;t) &= \frac{\mathbb{E}\left\{N\psi_\theta(a,a^*;\boldsymbol{O})-\Psi_I(a,a^*;0)\right\}S(\boldsymbol{O})]}{\mathbb{E}[N]} - \frac{\theta_I(a,a^*)}{\mathbb{E}[N]}\mathbb{E}[(N-\mathbb{E}[N])S(\boldsymbol{O})] \\
&= \mathbb{E}\left[\left\{\frac{N\psi_\theta(a,a^*;\boldsymbol{O})-\Psi_I(a,a^*;0)-N\theta_I(a,a^*)+\theta_I(a,a^*)\mathbb{E}[N]}{\mathbb{E}[N]}\right\}S(\boldsymbol{O})\right] \\
&= \mathbb{E}\left[\left\{\frac{N\psi_\theta(a,a^*;\boldsymbol{O})-N\theta_I(a,a^*)}{\mathbb{E}[N]}\right\}S(\boldsymbol{O})\right],
\end{aligned}$$

thus the EIF of $\theta_I(a,a^*;t)$ is $D_{\theta_I(a,a^*)}(\boldsymbol{O}) = \frac{N\psi_\theta(a,a^*;\boldsymbol{O})-N\theta_I(a,a^*)}{\mathbb{E}[N]}$. The EIF of τ_I can be derived using a similar strategy. □

Proof of Proposition 1. We shall suppress the subscripts in the density functions f for ease of exposition. By Bayes' formula, we have

$$f(\boldsymbol{m}|a,\boldsymbol{c},n) = \frac{f(a|\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m},\boldsymbol{c},n)}{f(a,\boldsymbol{c},n)} = \frac{f(a|\boldsymbol{m},\boldsymbol{c},n)f(\boldsymbol{m}|\boldsymbol{c},n)}{f(a)},$$

which further suggests that

$$\begin{aligned}
w^{(1)}(a,a^*,\boldsymbol{m},\boldsymbol{c},n) &= \frac{\kappa(a^*,\boldsymbol{m},\boldsymbol{c},n)}{\kappa(a,\boldsymbol{m},\boldsymbol{c},n)} = \frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a,\boldsymbol{c},n)} = \frac{f(a^*|\boldsymbol{m},\boldsymbol{c},n)}{f(a|\boldsymbol{m},\boldsymbol{c},n)}\frac{f(a)}{f(a^*)} \\
&= \frac{s(a^*,\boldsymbol{m},\boldsymbol{c},n)}{s(a,\boldsymbol{m},\boldsymbol{c},n)} \times \frac{\pi^a(1-\pi)^a}{\pi^{a^*}(1-\pi)^{a^*}},
\end{aligned}$$

and

$$\begin{aligned}
w^{(2)}(a,a^*,a',\boldsymbol{m},\boldsymbol{c},n) &= \frac{\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)\kappa_{\cdot(-j)}(a^*,m_{\cdot j},\boldsymbol{c},n)}{\kappa(a',\boldsymbol{m},\boldsymbol{c},n)} = \frac{f(m_{\cdot j}|a,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a',\boldsymbol{c},n)} \\
&= \frac{f(m_{\cdot j}|a,\boldsymbol{c},n)f(\boldsymbol{m}_{\cdot(-j)}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a',\boldsymbol{c},n)} \\
&= \frac{f(m_{\cdot j}|a,\boldsymbol{c},n)}{f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)}\frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(\boldsymbol{m}|a',\boldsymbol{c},n)} \\
&= \frac{f(m_{\cdot j}|a,\boldsymbol{c},n)}{f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)}\frac{f(a^*|\boldsymbol{m},\boldsymbol{c},n)}{f(a'|\boldsymbol{m},\boldsymbol{c},n)}\frac{f(a')}{f(a^*)} \\
&= \frac{\kappa_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)}{\kappa^\star_{\cdot j}(a^*,\boldsymbol{m},\boldsymbol{c},n)} \times \frac{s(a^*,\boldsymbol{m},\boldsymbol{c},n)}{s(a',\boldsymbol{m},\boldsymbol{c},n)} \times \frac{\pi^{a'}(1-\pi)^{1-a'}}{\pi^{a^*}(1-\pi)^{1-a^*}}.
\end{aligned}$$

The reparameterization of $u^{(1)}_{\cdot j}(a,a^*,\boldsymbol{c},n)$ and $u^{(4)}_{\cdot j}(a,a^*,a',\boldsymbol{c},n)$ directly follows from the definition of conditional expectation. Finally, to prove the reparameterization of $u^{(2)}_{\cdot j}(a,a^*,m_{\cdot,j},\boldsymbol{c},n)$, we observe

$$\begin{aligned}
f(\boldsymbol{m}_{\cdot(-j)}|a^*,\boldsymbol{c},n) &= \frac{f(\boldsymbol{m}_{\cdot(-j)}|a^*,\boldsymbol{c},n)f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)}{f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)} = \frac{f(\boldsymbol{m}|a^*,\boldsymbol{c},n)}{f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)} \\
&= \frac{f(m_{\cdot j}|a^*,\boldsymbol{c},n)}{f(m_{\cdot j}|a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)}f(\boldsymbol{m}_{\cdot(-j)}|a^*,m_{\cdot j},\boldsymbol{c},n)
\end{aligned}$$

and therefore

$$
\begin{aligned}
& u_{\cdot j}^{(2)}(a, a^*, m_{\cdot,j}, \boldsymbol{c}, n) \\
= & \int_{\boldsymbol{m}_{\cdot(-j)}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n) f(\boldsymbol{m}_{\cdot(-j)}|a^*, \boldsymbol{c}, n) \mathrm{d}\boldsymbol{m}_{\cdot(-j)} \\
= & \int_{\boldsymbol{m}_{\cdot(-j)}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{c}, n) \frac{f(m_{\cdot j}|a^*, \boldsymbol{c}, n)}{f(m_{\cdot j}|a^*, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{c}, n)} f(\boldsymbol{m}_{\cdot(-j)}|a^*, m_{\cdot j}, \boldsymbol{c}, n) \mathrm{d}\boldsymbol{m}_{\cdot(-j)} \\
= & \mathbb{E}_{\boldsymbol{M}_{\cdot(-j)}} \left[\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) \frac{f(M_{\cdot j}|a^*, \boldsymbol{C}, N)}{f(M_{\cdot j}|a^*, \boldsymbol{M}_{\cdot(-j)}, \boldsymbol{C}, N)} \middle| A = a^*, M_{\cdot j} = m_{\cdot j}, \boldsymbol{C} = \boldsymbol{c}, N = n \right].
\end{aligned}
$$

This completes the proof. □

Before proceed to the proof of Theorems 3 and 4, we first provide two lemmas. To facilitate our presentation, we define $\psi_\theta^\delta(a, a^*; \boldsymbol{O}) = \delta(N)\psi_\theta(a, a^*; \boldsymbol{O})$ and $\psi_\tau^\delta(\boldsymbol{O}) = \delta(N)\psi_\theta(\boldsymbol{O})$, where $\delta(N)$ is a known weighting function of the cluster size N. Therefore, for any approach $d \in \{\text{eif}_1\text{-par}, \text{eif}_2\text{-par}, \text{eif}_1\text{-ml}, \text{eif}_2\text{-ml}\}$, estimators of $\theta_V(a, a^*)$ and τ_V (for both $V \in \{I, C\}$) can be expressed as

$$
\begin{aligned}
\widehat{\theta}_V^d(a, a^*) &= \frac{1}{K \times \overline{\delta}} \sum_{i=1}^K \widehat{\psi}_\theta^\delta(a, a^*; \boldsymbol{O}) = \frac{1}{K \times \overline{\delta}} \sum_{i=1}^K \delta(N) \widehat{\psi}_\theta(a, a^*; \boldsymbol{O}), \\
\widehat{\tau}_V^d &= \frac{1}{K \times \overline{\delta}} \sum_{i=1}^K \widehat{\psi}_\tau^\delta(\boldsymbol{O}) = \frac{1}{K \times \overline{\delta}} \sum_{i=1}^K \delta(N) \widehat{\psi}_\tau(\boldsymbol{O}),
\end{aligned}
$$

where $\overline{\delta} = \frac{1}{K}\sum_{i=1}^K \delta(N_i)$. Here, we set $\delta(N) = 1$ and N if we consider cluster- or individual-average estimand respectively, evaluate $\{\widehat{\psi}_\theta^\delta(a, a^*; \boldsymbol{O}), \widehat{\psi}_\tau^\delta(\boldsymbol{O})\}$ based on its original parameterization if $d \in \{\text{eif}_1\text{-par}, \text{eif}_1\text{-ml}\}$ or its reparameterization in Proposition 1 if $d \in \{\text{eif}_2\text{-par}, \text{eif}_2\text{-ml}\}$, and use parametric method for the nuisance functions if $d \in \{\text{eif}_1\text{-par}, \text{eif}_2\text{-par}\}$ and machine learning method under cross-fitting for the nuisance functions if $d \in \{\text{eif}_1\text{-ml}, \text{eif}_2\text{-ml}\}$.

Lemma 1. *(Double robustness of the original EIFs) Let $\widehat{\psi}_\theta^\delta(a, a^*; \boldsymbol{O})$ and $\widehat{\psi}_\tau^\delta(\boldsymbol{O})$ be estimates of $\psi_\theta^\delta(a, a^*; \boldsymbol{O})$ and $\psi_\tau^\delta(\boldsymbol{O})$ according to the original parameterization based on estimates of the nuisance functions, $\widehat{h}_{nuisance}^{(1)} = \{\widehat{\kappa}_{\cdot j}, \widehat{\eta}_{\cdot j}, \widehat{\kappa}^c, \widehat{\kappa}^c_{\cdot(-j)}\}$. Then, we have that*

$$
\begin{aligned}
& \mathbb{E}[\widehat{\psi}_\theta^\delta(a, a^*; \boldsymbol{O}) - \psi_\theta^\delta(a, a^*; \boldsymbol{O})] \\
= & \mathbb{E}\left[\frac{\delta(N)}{N} \sum_{j=1}^N \int_{\boldsymbol{m}} \frac{\widehat{\kappa}^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)}{\widehat{\kappa}^c(a, \boldsymbol{m}, \boldsymbol{C}, N)} \left\{ \widehat{\kappa}^c(a, \boldsymbol{m}, \boldsymbol{C}, N) - \kappa^c(a, \boldsymbol{m}, \boldsymbol{C}, N) \right\} \left\{ \widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N) \right\} d\boldsymbol{m} \right] \\
& - \mathbb{E}\left[\frac{\delta(N)}{N} \sum_{j=1}^N \int_{\boldsymbol{m}} \left\{ \widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N) \right\} \left\{ \widehat{\kappa}^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N) - \kappa^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N) \right\} d\boldsymbol{m} \right]
\end{aligned}
$$

and

$$
\mathbb{E}[\widehat{\psi}_\tau^\delta(\boldsymbol{O}) - \psi_\tau^\delta(\boldsymbol{O})]
$$

$$=\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\frac{\widehat{\kappa}_{\cdot j}(1,m_j,\boldsymbol{C},N)\widehat{\kappa}^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)}{\widehat{\kappa}^c(1,\boldsymbol{m},\boldsymbol{C},N)}\left\{\widehat{\kappa}^c(1,\boldsymbol{m},\boldsymbol{C},N)-\kappa^c(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}d\boldsymbol{m}\right]$$

$$-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\widehat{\kappa}^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}d\boldsymbol{m}\right]$$

$$-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)-\kappa^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\right\}d\boldsymbol{m}\right]$$

$$-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\left\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)-\kappa^c_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\right\}d\boldsymbol{m}\right].$$

Therefore, we have that (i) $\widehat{\psi}^{\delta}_{\theta}(a,a^*;\boldsymbol{O})$ is unbiased to $\psi^{\delta}_{\theta}(a,a^*;\boldsymbol{O})$ when either $\widehat{\eta}_{\cdot j}$ or $\widehat{\kappa}^c$ converges to its true value and (ii) $\widehat{\psi}^{\delta}_{\tau}(\boldsymbol{O})$ is unbiased to $\psi^{\delta}_{\tau}(\boldsymbol{O})$ when either $\{\widehat{\eta}_{\cdot j},\widehat{\kappa}_{\cdot j}\}$ or $\{\widehat{\kappa}_{\cdot j},\widehat{\kappa}^c_{\cdot(-j)},\widehat{\kappa}^c\}$ converge to their true values.

Proof of Lemma 1. We can show that

$$\begin{aligned}
&\mathbb{E}[\widehat{\psi}^{\delta}_{\theta}(a,a^*;\boldsymbol{O})-\psi^{\delta}_{\theta}(a,a^*;\boldsymbol{O})]\\
=&\mathbb{E}\Bigg[\frac{\delta(N)}{N}\sum_{j=1}^{N}\Bigg\{\frac{\mathbb{I}(A=a)}{\pi^a(1-\pi)^{1-a}}\widehat{w}^{(1)}(a,a^*,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j}-\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\\
&+\frac{\mathbb{I}(A=a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)-\widehat{u}^{(1)}_{\cdot j}(a,a^*,\boldsymbol{C},N)\right\}+\widehat{u}^{(1)}_{\cdot j}(a,a^*,\boldsymbol{C},N)\Bigg\}\Bigg]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(1)}_{\cdot j}(a,a^*,\boldsymbol{C},N)\right]\\
=&\mathbb{E}\Bigg[\frac{\delta(N)}{N}\sum_{j=1}^{N}\Bigg\{\frac{f(A=a|\boldsymbol{M},\boldsymbol{C},N)}{f(A=a)}\widehat{w}^{(1)}(a,a^*,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j}-\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\Bigg]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=a^*|\boldsymbol{M},\boldsymbol{C},N)}{f(A=a^*)}\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(1)}_{\cdot j}(a,a^*,\boldsymbol{C},N)\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa^c(a,\boldsymbol{m},\boldsymbol{C},N)\frac{\widehat{\kappa}^c(a^*,\boldsymbol{m},\boldsymbol{C},N)}{\widehat{\kappa}^c(a,\boldsymbol{m},\boldsymbol{C},N)}\left\{\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa^c(a^*,\boldsymbol{m},\boldsymbol{C},N)\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa^c(a^*,\boldsymbol{m},\boldsymbol{C},N)\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa^c(a,\boldsymbol{m},\boldsymbol{C},N)\frac{\widehat{\kappa}^c(a^*,\boldsymbol{m},\boldsymbol{C},N)}{\widehat{\kappa}^c(a,\boldsymbol{m},\boldsymbol{C},N)}\left\{\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\kappa^c(a^*,\boldsymbol{m},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\frac{\widehat{\kappa}^c(a^*,\boldsymbol{m},\boldsymbol{C},N)}{\widehat{\kappa}^c(a,\boldsymbol{m},\boldsymbol{C},N)}\left\{\widehat{\kappa}^c(a,\boldsymbol{m},\boldsymbol{C},N)-\kappa^c(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}^c(a^*,\boldsymbol{m},\boldsymbol{C},N)-\kappa^c(a^*,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right].
\end{aligned}$$

Also, we have that

$$
\begin{aligned}
&\mathbb{E}[\widehat{\psi}_{\tau}^{\delta}(\boldsymbol{O}) - \psi_{\tau}^{\delta}(\boldsymbol{O})] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{A}{\pi}\widehat{w}^{(2)}(1,0,1,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j} - \widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{A}{\pi}\left\{\widehat{u}_{\cdot j}^{(2)}(1,0,M_{\cdot j},\boldsymbol{C},N) - \widehat{u}_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{1-A}{1-\pi}\left\{\widehat{u}_{\cdot j}^{(3)}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N) - \widehat{u}_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\widehat{u}_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right] - \mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=1|\boldsymbol{M},\boldsymbol{C},N)}{f(A=1)}\widehat{w}^{(2)}(1,0,1,\boldsymbol{M},\boldsymbol{C},N)\left\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N) - \widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=1|M_{\cdot j},\boldsymbol{C},N)}{f(A=1)}\widehat{u}_{\cdot j}^{(2)}(1,0,M_{\cdot j},\boldsymbol{C},N)\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=0|\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)}{f(A=0)}\widehat{u}_{\cdot j}^{(3)}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\widehat{u}_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right] - \mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u_{\cdot j}^{(4)}(1,1,0,\boldsymbol{C},N)\right] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa^{c}(1,\boldsymbol{m},\boldsymbol{C},N)\frac{\widehat{\kappa}_{\cdot j}(1,m_{j},\boldsymbol{C},N)\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)}{\widehat{\kappa}(1,\boldsymbol{m},\boldsymbol{C},N)}\left\{\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N) - \widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\frac{\widehat{\kappa}_{\cdot j}(1,m_{j},\boldsymbol{C},N)\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)}{\widehat{\kappa}^{c}(1,\boldsymbol{m},\boldsymbol{C},N)}\left\{\widehat{\kappa}^{c}(1,\boldsymbol{m},\boldsymbol{C},N) - \kappa^{c}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N) - \eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N) - \eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N) - \kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\left\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N) - \eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N) - \kappa_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\left\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N) - \kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\left\{\widehat{\kappa}_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N) - \kappa_{\cdot(-j)}^{c}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\right\}\mathrm{d}\boldsymbol{m}\right]
\end{aligned}
$$

□

Lemma 2. *(Double robustness of the EIFs after reparameterization of Proposition 1) Let* $\widehat{\psi}_{\theta}^{\delta}(a,a^{*};\boldsymbol{O})$ *and* $\widehat{\psi}_{\tau}^{\delta}(\boldsymbol{O})$ *be estimates of* $\psi_{\theta}^{\delta}(a,a^{*};\boldsymbol{O})$ *and* $\psi_{\tau}^{\delta}(\boldsymbol{O})$ *according to their repa-*

rameterization in Propostion 1 based on estimates of the nuisance functions, $\widehat{h}^{(2)}_{nuisance} = \{\widehat{\eta}_{\cdot j}, \widehat{\eta}^{\star}_{\cdot j}, \widehat{\eta}^{\dagger}_{\cdot j}, \widehat{\kappa}_{\cdot j}, \widehat{\kappa}^{\star}_{\cdot j}, \widehat{s}\}$*. Then, we have that*

$$
\begin{aligned}
&\mathbb{E}[\widehat{\psi}^{\delta}_{\theta}(a, a^*; \boldsymbol{O}) - \psi^{\delta}_{\theta}(a, a^*; \boldsymbol{O})] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{\widehat{s}(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=a^*)}\left\{\widehat{s}(a, \boldsymbol{M}, \boldsymbol{C}, N) - s(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\left\{\widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) - \eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{1}{f(A=a^*)}\left\{\widehat{s}(a^*, \boldsymbol{M}, \boldsymbol{C}, N) - s(a^*, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\left\{\widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) - \eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[\widehat{\psi}^{\delta}_{\tau}(\boldsymbol{O}) - \psi^{\delta}_{\tau}(\boldsymbol{O})] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(1, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=0)}\left\{\frac{\widehat{\kappa}_{\cdot j}(1, M_{\cdot j}, \boldsymbol{C}, n)}{\widehat{\kappa}^{\star}_{\cdot j}(0, \boldsymbol{M}, \boldsymbol{C}, N)}\frac{\widehat{s}(0, \boldsymbol{M}, \boldsymbol{C}, N)}{\widehat{s}(1, \boldsymbol{M}, \boldsymbol{C}, N)} - \frac{\kappa_{\cdot j}(1, M_{\cdot j}, \boldsymbol{C}, n)}{\kappa^{\star}_{\cdot j}(0, \boldsymbol{M}, \boldsymbol{C}, N)}\frac{s(0, \boldsymbol{M}, \boldsymbol{C}, N)}{s(1, \boldsymbol{M}, \boldsymbol{C}, N)}\right\}\left\{\eta_{\cdot j}(1, \boldsymbol{M}, \boldsymbol{C}, N) - \widehat{\eta}_{\cdot j}(1, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\left\{\widehat{u}^{(2)}_{\cdot j}(1, 0, m_{\cdot j}, \boldsymbol{C}, N) - u^{(2)}_{\cdot j}(1, 0, m_{\cdot j}, \boldsymbol{C}, N)\right\}\left\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\right\}dm_{\cdot j}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\left\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\right\}\left\{\widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\right\}d\boldsymbol{m}\right].
\end{aligned}
$$

Therefore, we have that (i) $\widehat{\psi}_{\theta}(a, a^*; \boldsymbol{O})$ *is unbiased to* $\psi_{\theta}(a, a^*; \boldsymbol{O})$ *when either* $\widehat{s}$ *or* $\widehat{\eta}_{\cdot j}$ *converges to its true value and (ii)* $\widehat{\psi}_{\tau}(\boldsymbol{O})$ *is unbiased to* $\psi_{\tau}(\boldsymbol{O})$ *when either* $\{\widehat{\eta}_{\cdot j}, \widehat{\kappa}_{\cdot j}\}$ *or* $\{\widehat{\kappa}_{\cdot j}, \widehat{\kappa}^{\star}_{\cdot j}, \widehat{s}\}$ *converge to their true values.*

Proof of Lemma 2. We can show that

$$
\begin{aligned}
&\mathbb{E}[\widehat{\psi}^{\delta}_{\theta}(a, a^*; \boldsymbol{O}) - \psi^{\delta}_{\theta}(a, a^*; \boldsymbol{O})] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{\mathbb{I}(A=a)}{\pi^{a}(1-\pi)^{1-a}}\widehat{w}^{(1)}(a, a^*, \boldsymbol{M}, \boldsymbol{C}, N)\left\{Y_{\cdot j} - \widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{\mathbb{I}(A=a^*)}{\pi^{a^*}(1-\pi)^{1-a^*}}\left\{\widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) - \widehat{u}^{(1)}_{\cdot j}(a, a^*, \boldsymbol{C}, N)\right\} + \widehat{u}^{(1)}_{\cdot j}(a, a^*, \boldsymbol{C}, N)\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(1)}_{\cdot j}(a, a^*, \boldsymbol{C}, N)\right] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(a, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=a)}\frac{\widehat{s}(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}{\widehat{s}(a, \boldsymbol{M}, \boldsymbol{C}, N)}\frac{f(A=a)}{f(A=a^*)}\left\{\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) - \widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=a^*)}\widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right] \\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(1)}_{\cdot j}(a, a^*, \boldsymbol{C}, N)\right] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(a, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=a)}\frac{\widehat{s}(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}{\widehat{s}(a, \boldsymbol{M}, \boldsymbol{C}, N)}\frac{f(A=a)}{f(A=a^*)}\left\{\eta_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N) - \widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right\}\right] \\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(a^*, \boldsymbol{M}, \boldsymbol{C}, N)}{f(A=a^*)}\widehat{\eta}_{\cdot j}(a, \boldsymbol{M}, \boldsymbol{C}, N)\right]
\end{aligned}
$$

$$
\begin{aligned}
&- \mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(a^*,\boldsymbol{M},\boldsymbol{C},N)}{f(A=a^*)}\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{\widehat{s}(a^*,\boldsymbol{M},\boldsymbol{C},N)}{f(A=a^*)}\left\{\widehat{s}(a,\boldsymbol{M},\boldsymbol{C},N)-s(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)-\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]\\
&- \mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{1}{f(A=a^*)}\left\{\widehat{s}(a^*,\boldsymbol{M},\boldsymbol{C},N)-s(a^*,\boldsymbol{M},\boldsymbol{C},N)\right\}\left\{\widehat{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)-\eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]
\end{aligned}
$$

Also, we have that

$$
\begin{aligned}
&\mathbb{E}[\widehat{\psi}^{\delta}_{\tau}(\boldsymbol{O})-\psi^{\delta}_{\tau}(\boldsymbol{O})]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{A}{\pi}\widehat{w}^{(2)}(1,0,1,\boldsymbol{M},\boldsymbol{C},N)\left\{Y_{\cdot j}-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{A}{\pi}\left\{\widehat{u}^{(2)}_{\cdot j}(1,0,M_{\cdot j},\boldsymbol{C},N)-\widehat{u}^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right\}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{1-A}{1-\pi}\left\{\widehat{u}^{(3)}_{\cdot j}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)-\widehat{u}^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right\}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\widehat{u}^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right]-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=1|\boldsymbol{M},\boldsymbol{C},N)}{f(A=1)}\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},N)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}\frac{f(A=1)}{f(A=0)}\left\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=1|M_{\cdot j},\boldsymbol{C},N)}{f(A=1)}\widehat{u}^{(2)}_{\cdot j}(1,0,M_{\cdot j},\boldsymbol{C},N)\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{f(A=0|\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)}{f(A=0)}\widehat{u}^{(3)}_{\cdot j}(1,1,\boldsymbol{M}_{\cdot(-j)},\boldsymbol{C},N)\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\widehat{u}^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right]-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}u^{(4)}_{\cdot j}(1,1,0,\boldsymbol{C},N)\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(1,\boldsymbol{M},\boldsymbol{C},N)}{f(A=0)}\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}\left\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{u}^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}m_{\cdot j}\right]\\
&+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}_{\cdot(-j)}}\int_{m_{\cdot j}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}m_{\cdot j}\times\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}_{\cdot(-j)}\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{u}^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}m_{\cdot j}\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]\\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(1,\boldsymbol{M},\boldsymbol{C},N)}{f(A=0)}\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}\left\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\right\}\right]\\
&-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{u}^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\left\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\mathrm{d}m_{\cdot j}\right]
\end{aligned}
$$

$$+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\left\{\int_{\boldsymbol{m}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]$$

$$=\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(1,\boldsymbol{M},\boldsymbol{C},N)}{f(A=0)}\left\{\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}-\frac{\kappa_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\kappa^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{s(0,\boldsymbol{M},\boldsymbol{C},N)}{s(1,\boldsymbol{M},\boldsymbol{C},N)}\right\}\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\}\right]$$

$$+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\{\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\}\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]$$

$$-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{u}^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}m_{\cdot j}\right]$$

$$+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\left\{\int_{\boldsymbol{m}}\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\right\}\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\mathrm{d}\boldsymbol{m}\right]$$

$$=\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\frac{s(1,\boldsymbol{M},\boldsymbol{C},N)}{f(A=0)}\left\{\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}-\frac{\kappa_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\kappa^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{s(0,\boldsymbol{M},\boldsymbol{C},N)}{s(1,\boldsymbol{M},\boldsymbol{C},N)}\right\}\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\}\right]$$

$$-\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\left\{\widehat{u}^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)-u^{(2)}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\right\}\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}m_{\cdot j}\right]$$

$$+\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\}\,\mathrm{d}\boldsymbol{m}\right]$$

$\square$

Proof of Theorem 3. For our proposed efficient estimator with user-specified parametric working models, let $\Psi^{\text{par}}_{\zeta}(\boldsymbol{O};\zeta,\boldsymbol{\alpha})$ be the estimating function of $\zeta\in\{\theta_C(a,a^*),\theta_I(a,a^*),\tau_C,\tau_I\}$, where $\Psi^{\text{par}}_{\zeta}(\boldsymbol{O};\zeta,\boldsymbol{\alpha})$ are specified as the EIF in Theorem 2 and $\boldsymbol{\alpha}$ is finite-dimensional nuisance parameters depending on the specified parametric working models. Therefore, both $\widehat{\zeta}^{\text{eif}_1\text{-par}}$ and $\widehat{\zeta}^{\text{eif}_2\text{-par}}$ can be seen as the solution of $\mathbb{P}_K[\Psi^{\text{par}}_{\zeta}(\boldsymbol{O};\zeta,\widehat{\boldsymbol{\alpha}})]=0$, where the difference lies that the first estimator uses the original EIF and the second estimator uses the EIF after the reparameterization given in Proposition 1. When the original EIFs are used, one can decompose $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_{\eta_{\cdot j}},\boldsymbol{\alpha}_{\kappa_{\cdot j}},\boldsymbol{\alpha}_{\mathcal{C}})$ as unknown parameters in the working models of $\{\eta_{\cdot j},\kappa_{\cdot j},\mathcal{C}\}$ that are used to estimate $h^{(1)}_{\text{nuisance}}=\{\eta_{\cdot j},\kappa_{\cdot j},\kappa^c,\kappa^c_{\cdot(-j)}\}$. If the reparameterization in Proposition 1 is used, one can similarly write out $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_{\eta_{\cdot j}},\boldsymbol{\alpha}_{\eta^{\star}_{\cdot j}},\boldsymbol{\alpha}_{\eta^{\dagger}_{\cdot j}},\boldsymbol{\alpha}_{\kappa_{\cdot j}},\boldsymbol{\alpha}_{\kappa^{\star}_{\cdot j}},\boldsymbol{\alpha}_{s})$ as unknown parameters in the working models of $h^{(2)}_{\text{nuisance}}=\{\eta_{\cdot j},\eta^{\star}_{\cdot j},\eta^{\dagger}_{\cdot j},\kappa_{\cdot j},\kappa^{\star}_{\cdot j},s\}$, respectively. Specific expressions of $\Psi^{\text{par}}_{\zeta}(\boldsymbol{O};\zeta,\boldsymbol{\alpha})$ are

$$\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\theta_V(a,a^*),\boldsymbol{\alpha})=\psi^{\delta}_{\theta}(a,a^*;\boldsymbol{O};\widehat{\boldsymbol{\alpha}})-\delta(N)\times\theta_V(a,a^*),$$
$$\Psi^{\text{par}}_{\tau_V}(\boldsymbol{O};\tau_V,\boldsymbol{\alpha})=\psi^{\delta}_{\tau}(\boldsymbol{O};\widehat{\boldsymbol{\alpha}})-\delta(N)\times\tau_V,$$

where $\delta(N)=1$ and N if cluster-average ($V=C$) or individual-average ($V=I$) estimands are considered, respectively, and $\{\psi^{\delta}_{\theta}(a,a^*;\boldsymbol{O};\widehat{\boldsymbol{\alpha}}),\psi^{\delta}_{\tau}(\boldsymbol{O};\widehat{\boldsymbol{\alpha}})\}$ are values of $\{\psi^{\delta}_{\theta}(a,a^*;\boldsymbol{O}),\psi^{\delta}_{\tau}(\boldsymbol{O})\}$ evaluated based on parametric working models. We assume the following regularity conditions:

1. Assume $\boldsymbol{\alpha}\in\mathcal{A}$, where the parameter space $\mathcal{A}$ is a compact subset of the Euclidean space. Suppose that $\widehat{\boldsymbol{\alpha}}$ converges to an interior point in $\mathcal{A}$, denoted by $\underline{\boldsymbol{\alpha}}$. In addition, we assume that $\widehat{\boldsymbol{\alpha}}$ is $K^{-1/2}$-consistent to $\underline{\boldsymbol{\alpha}}$ and asymptotically linear with

the influence function $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})$ such that $\sqrt{K}(\widehat{\boldsymbol{\alpha}} - \underline{\boldsymbol{\alpha}}) = \sqrt{K}\mathbb{P}_K[\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})] + o_p(1)$ with $\mathbb{E}[\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})] = \mathbf{0}$ and $\mathbb{E}[\Psi_{\boldsymbol{\alpha}}^{\otimes 2}(\boldsymbol{O})] < \infty$, where $\boldsymbol{X}^{\otimes 2} = \boldsymbol{X}\boldsymbol{X}^T$ for a vector $\boldsymbol{X}$.

2. Let ξ be a nuisance function in $h^{(1)}_{\text{nuisance}}$ or $h^{(2)}_{\text{nuisance}}$, and $\underline{\xi}$ be the value of ξ when it is evaluated based on the parametric working model at $\boldsymbol{\alpha}_\xi = \underline{\boldsymbol{\alpha}}_\xi$. For every ξ, we assume $\underline{\xi}$ equals to the true value of ξ if the corresponding parametric model is correctly specified.

3. The true value of ζ, denoted by $\underline{\zeta}$, is considered as the unique solution of $\mathbb{E}[\Psi_\zeta(\boldsymbol{O};\zeta)] = 0$, where $\Psi_\zeta(\boldsymbol{O};\zeta)$ is the EIF-induced estimating function of ζ where all nuisance functions in the EIF are evaluated at their true value.

4. Suppose that $(\zeta, \boldsymbol{\alpha}) \mapsto \Psi^{\text{par}}_\zeta(\boldsymbol{O};\zeta,\boldsymbol{\alpha})$ is P-Donsker, dominated by a square-integrable function, and also twice continuously differentiable in a small neighborhood around $(\underline{\zeta}, \underline{\boldsymbol{\alpha}})$.

Next we prove the consistency, asymptotic normality, and local efficiency of $\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*)$, where proofs for other estimators are similar and therefore are omitted. Specifically, if either $\eta_{\cdot j}$ or $\{\kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified, we can show that

$$
\begin{aligned}
&\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\theta_V(a,a^*),\underline{\boldsymbol{\alpha}}) - \Psi_{\theta_V(a,a^*)}(\boldsymbol{O};\theta_V(a,a^*))] \\
=&\mathbb{E}[\delta(N)\psi_\theta(a,a^*;\boldsymbol{O};\underline{\boldsymbol{\alpha}}) - \delta(N)\theta_C(a,a^*) - \delta(N)\psi_\theta(a,a^*;\boldsymbol{O}) + \delta(N)\theta_C(a,a^*)] \\
=&\mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^N \frac{\underline{s}(a^*,\boldsymbol{M},\boldsymbol{C},N)}{f(A=a^*)}\{\underline{s}(a,\boldsymbol{M},\boldsymbol{C},N) - s(a,\boldsymbol{M},\boldsymbol{C},N)\}\left\{\underline{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N) - \eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\right] \\
&- \mathbb{E}\left[\frac{\delta(N)}{N}\sum_{j=1}^N \frac{1}{f(A=a^*)}\{\underline{s}(a^*,\boldsymbol{M},\boldsymbol{C},N) - s(a^*,\boldsymbol{M},\boldsymbol{C},N)\}\left\{\underline{\eta}_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N) - \eta_{\cdot j}(a,\boldsymbol{M},\boldsymbol{C},N)\right\}\right] \\
&= 0, \qquad \text{(s18)}
\end{aligned}
$$

where the first equality holds by the definition of the EIF of $\theta_V(a,a^*)$, the second equality holds by the doubly robust property in Lemma 1, and the third equality holds by regularity condition 2 under the scenario that either $\eta_{\cdot j}$ or $\{\kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified. Literally, the above discussion suggests that $\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\theta_V(a,a^*),\widehat{\boldsymbol{\alpha}})$ is an unbiased estimating function of $\theta_V(a,a^*)$ if either $\eta_{\cdot j}$ or $\{\kappa_{\cdot j}, \mathcal{C}\}$ are correctly specified. To proceed, we use a Taylor series, along with regularity condition 4, to deduce that

$$
\begin{aligned}
0 =&\mathbb{P}_K[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*),\widehat{\boldsymbol{\alpha}})] \\
=&\mathbb{P}_K[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})] + \frac{\partial\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})]}{\partial\underline{\theta}_V(a,a^*)}\left\{\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*) - \underline{\theta}_V(a,a^*)\right\} \\
&+ \frac{\partial\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})]}{\partial\underline{\boldsymbol{\alpha}}}\{\widehat{\boldsymbol{\alpha}} - \underline{\boldsymbol{\alpha}}\} + o_p(K^{-1/2}) \\
=&\mathbb{P}_K[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})] - \mathbb{E}[\delta(N)]\left\{\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*) - \underline{\theta}_V(a,a^*)\right\} + \boldsymbol{B}_{\boldsymbol{\alpha}}\{\widehat{\boldsymbol{\alpha}} - \underline{\boldsymbol{\alpha}}\} + o_p(K^{-1/2}),
\end{aligned}
$$

where the partial derivative $\boldsymbol{B}_{\boldsymbol{\alpha}} := \frac{\partial\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})]}{\partial\underline{\boldsymbol{\alpha}}}$ exists due to regularity condition

4. This further implies that

$$\begin{aligned}\sqrt{K}\left\{\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*)-\underline{\theta}_V(a,a^*)\right\} &= \frac{\sqrt{K}}{\mathbb{E}[\delta(N)]}\mathbb{P}_K[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})]+\frac{\sqrt{K}}{\mathbb{E}[\delta(N)]}\mathbb{P}_K[\boldsymbol{B}_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})]+o_p(1)\\ &= \frac{\sqrt{K}}{\mathbb{E}[\delta(N)]}\mathbb{P}_K[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})+\boldsymbol{B}_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})]+o_p(1),\end{aligned}$$

where the first equality follows regularity condition 1. Noting that (i) $\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})]=0$ due to (s18) along with regularity condition 3 and (ii) $\mathbb{E}[\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})]=\boldsymbol{0}$, we further obtain

$$\mathbb{E}[\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})+\boldsymbol{B}_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})]=0.$$

In addition, due to regularity conditions 1 and 4, one can show

$$\mathbb{E}\left[\left\{\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})+\boldsymbol{B}_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})\right\}^2\right]<\infty.$$

Then, following the central limit theorem, we deduce that $\sqrt{K}\left\{\widehat{\theta}_V^{\text{eif}_1\text{-par}}(a,a^*)-\underline{\theta}_V(a,a^*)\right\}$ converges to a normal distribution with mean 0 and variance

$$\Sigma^{\text{eif}_1\text{-par}}_{\theta_V(a,a^*)}=\mathbb{E}\left[\left\{\frac{\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})+\boldsymbol{B}_{\boldsymbol{\alpha}}\Psi_{\boldsymbol{\alpha}}(\boldsymbol{O})}{\mathbb{E}[\delta(N)]}\right\}^2\right]$$

Finally, when all parametric working models are correctly specified, we have that $\Sigma^{\text{eif}_1\text{-par}}_{\theta_V(a,a^*)}=\mathbb{E}[\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})^2]$ achieves the semiparametric efficiency bound, because (i) $\boldsymbol{B}_{\boldsymbol{\alpha}}=\boldsymbol{0}$ due to the orthogonality of EIF and (ii) $\Psi^{\text{par}}_{\theta_V(a,a^*)}(\boldsymbol{O};\underline{\theta}_V(a,a^*),\underline{\boldsymbol{\alpha}})/\mathbb{E}[\delta(N)]=\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})$ when all working models are correct.

Following the same strategy, one can conclude the robustness and efficiency properties among other estimators. In the proof of $\widehat{\zeta}^{\text{eif}_2\text{-par}}$, one should remind that one need to use the double robustness property in Lemma 2 to prove the unbiasedness of EIF-induced estimating functions in (s18).

□

Proof of Theorem 4. Now we consider using machine learners or nonparametric methods to estimate $\widehat{h}^{(1)}_{nuisance}$ and $\widehat{h}^{(2)}_{nuisance}$ according to the the cross-fitting procedure detailed in Section 4.3. We assume the following regularity conditions regarding $\widehat{h}^{(1)}_{nuisance}$ and $\widehat{h}^{(2)}_{nuisance}$:

1. For $\widehat{h}^{(1)}_{nuisance}=\{\widehat{\eta}_{\cdot j},\widehat{\kappa}_{\cdot j},\widehat{\kappa}^c,\widehat{\kappa}^c_{\cdot(-j)}\}$, assume that (i) $\widehat{\kappa}_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)$, $\widehat{\kappa}^c(a,\boldsymbol{m},\boldsymbol{c},n)$, and $\widehat{\kappa}^c_{\cdot(-j)}(a,m_{\cdot(-j)},\boldsymbol{c},n)$ are bounded away from 0 for $\{a,\boldsymbol{m},\boldsymbol{c},n\}$ over their valid support, and (ii) $\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$ is uniformly bounded for $\{a,\boldsymbol{m},\boldsymbol{c},n\}$ over their valid support.

2. For $\widehat{h}^{(2)}_{\text{nuisance}}=\{\widehat{\eta}_{\cdot j},\widehat{\eta}^{\star}_{\cdot j},\widehat{\eta}^{\dagger}_{\cdot j},\widehat{\kappa}_{\cdot j},\widehat{\kappa}^{\star}_{\cdot j},\widehat{s}\}$, assume that (i) $\widehat{\kappa}_{\cdot j}(a,m_{\cdot j},\boldsymbol{c},n)$, $\widehat{\kappa}^{\star}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$, $\widehat{s}(a,\boldsymbol{m},\boldsymbol{c},n)$ are are bounded away from 0 for $\{a,\boldsymbol{m},\boldsymbol{c},n\}$ over their valid support, and (ii) $\widehat{\eta}_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)$, $\widehat{\eta}^{\star}_{\cdot j}(a,a^*,\boldsymbol{c},n)$, $\widehat{\eta}^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$ are uniformly bounded for $\{a,a^*,\boldsymbol{m},\boldsymbol{c},n\}$ over their valid support.

Next, we derive the asymptotic property of $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$, where the asymptotic properties among other machine learning estimators can be similarly obtained and are omitted here. We shall use $\underline{\theta}_V(a,a^*)$ to denote the true value of $\theta_V(a,a^*)$, which equals to

$\mathbb{E}[\psi_\theta^\delta(a,a^*;\boldsymbol{O})]/\mathbb{E}[\delta(N)]$ by Theorem 2, where $\delta(N)=1$ or N if we consider the cluster-average or individual-average estimand, respectively. Our central task is to show that $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$ converges to $\underline{\theta}_V(a,a^*)$ if $\widehat{\kappa}^c$ or $\widehat{\eta}_{\cdot j}$ is consistent in $L_2(P)$-norm, and $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$ is asymptotically normal and locally efficient when both $\{\widehat{\kappa}^c,\widehat{\eta}_{\cdot j}\}$ are consistent with $\lambda_K(\widehat{\kappa}^c)\lambda_K(\widehat{\eta}_{\cdot j})=o(K^{-1/2})$. Consistency of $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$ follows directly from Lemma 1 by noting that $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$ converges in probability to $\mathbb{E}[\widehat{\psi}_\theta^\delta(a,a^*;\boldsymbol{O})]/\mathbb{E}[\delta(N)]$ and $\mathbb{E}[\widehat{\psi}_\theta^\delta(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]=o_p(1)$ by Lemma 1 along with the condition that either $\widehat{\kappa}^c$ or $\widehat{\eta}_{\cdot j}$ is consistent in $L_2(P)$-norm.

The proof of asymptotic normality and local efficiency is more involved. Based on cross-fitting, we can rewrite $\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)$ as

$$\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a,a^*)=\frac{1}{K\times\bar{\delta}}\sum_{r=1}^R K_r\mathbb{P}_{K_r}[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})],$$

where K_r is the sample size in the r-th group $\boldsymbol{\mathcal{S}}_r$, $\mathbb{P}_{K_r}[\cdot]$ is the the empirical mean operator on $\boldsymbol{\mathcal{S}}_r$, and $\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})$ is the estimate of $\psi_\theta^\delta(a,a^*;\boldsymbol{O})$ with nuisance functions $\widehat{h}^{(1)}_{nuisance}$ training based on $\boldsymbol{\mathcal{S}}_{-r}$. We further have the following decomposition of $\mathbb{P}_{K_r}[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})]$:

$$\begin{aligned}\mathbb{P}_{K_r}[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})]=&\mathbb{P}_{K_r}[\psi_\theta^\delta(a,a^*;\boldsymbol{O})]+\underbrace{(\mathbb{P}_{K_r}-\mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]}_{=:R_1(\widehat{\psi}_\theta^{\delta,r},\psi_\theta^\delta)}\\&+\underbrace{\mathbb{E}[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]}_{=:R_2(\widehat{\psi}_\theta^{\delta,r},\psi_\theta^\delta)}.\end{aligned}$$

We can show $R_1(\widehat{\psi}_\theta^{\delta,r},\psi_\theta^\delta)$ is $o_p(K^{-1/2})$ if both $\{\widehat{\eta}_{\cdot j},\widehat{\kappa}^c\}$ are consistent under cross-fitting. Specifically, by the independence induced by cross-fitting, we have that

$$\begin{aligned}&\text{Var}\left\{(\mathbb{P}_{K_r}-\mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]\Big|\boldsymbol{\mathcal{S}}_{-r}\right\}\\=&\text{Var}\left\{\mathbb{P}_{K_r}[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]\Big|\boldsymbol{\mathcal{S}}_{-r}\right\}\\=&\frac{1}{K_r}\text{Var}\left\{\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})\Big|\boldsymbol{\mathcal{S}}_{-r}\right\}\\=&\frac{1}{K_r}\|\widehat{\psi}_\theta^{\delta,r}-\psi_\theta^\delta\|^2.\end{aligned}$$

This further suggests

$$\begin{aligned}&\mathbb{P}\left\{\frac{\sqrt{K_r}\left|(\mathbb{P}_{K_r}-\mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]\right|}{\|\widehat{\psi}_\theta^{\delta,r}-\psi_\theta^\delta\|}\geq\epsilon\right\}\\=&\mathbb{E}\left[\mathbb{P}\left\{\frac{\sqrt{K_r}\left|(\mathbb{P}_{K_r}-\mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]\right|}{\|\widehat{\psi}_\theta^{\delta,r}-\psi_\theta^\delta\|}\geq\epsilon\Big|\boldsymbol{\mathcal{S}}_{-v}\right\}\right]\\\leq&\frac{1}{\epsilon^2}\mathbb{E}\left[\text{Var}\left\{\frac{\sqrt{K_r}\left|(\mathbb{P}_{K_r}-\mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a,a^*;\boldsymbol{O})-\psi_\theta^\delta(a,a^*;\boldsymbol{O})]\right|}{\|\widehat{\psi}_\theta^{\delta,r}-\psi_\theta^\delta\|}\Big|\boldsymbol{\mathcal{S}}_{-v}\right\}\right]\end{aligned}$$

$$=\epsilon^{-2},$$

for any $\epsilon > 0$. Therefore, we have that $R_1(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta) = (\mathbb{P}_{K_r} - \mathbb{E})[\widehat{\psi}_\theta^{\delta,r}(a, a^*; \boldsymbol{O}) - \psi_\theta^\delta(a, a^*; \boldsymbol{O})] = O_p(K_r^{-1/2}\|\widehat{\psi}_\theta^{\delta,r} - \psi_\theta^\delta\|) = o_p(K_r^{-1/2})$. Since R is a finite number and we partition the dataset as evenly as possible, we have that $K_r/K = O(1)$ and thus $R_1(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta) = o_p(K^{-1/2})$. Next we analyze $R_2(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta)$. Specifically, we have

$$\begin{aligned} R_2(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta) =& \mathbb{E}[\widehat{\psi}_\theta^{\delta,r}(a, a^*; \boldsymbol{O}) - \psi_\theta^\delta(a, a^*; \boldsymbol{O})] \\ =& O_p(\|\widehat{\eta}_{\cdot j}^r - \eta_{\cdot j}\| \times \|\widehat{\kappa}^{c,r} - \kappa^c\|) \\ =& o_p(K^{-1/2}), \end{aligned}$$

where the second equality holds by Lemma 1 with the Cauchy-Schwarz inequality, the third equality holds because we assume the rate of convergence of the nuisance functions statisfying $\lambda_K(\widehat{\kappa}^c)\lambda_K(\widehat{\eta}_{\cdot j}) = o(K^{-1/2})$. We have now established $R_1(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta) = o_p(K^{-1/2})$ and $R_2(\widehat{\psi}_\theta^{\delta,r}, \psi_\theta^\delta) = o_p(K^{-1/2})$, which suggests that

$$\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a, a^*) = \frac{1}{K \times \overline{\delta}} \sum_{r=1}^R K_r \mathbb{P}_{K_r}[\widehat{\psi}_\theta^{\delta,r}(a, a^*; \boldsymbol{O})] = \frac{\mathbb{P}_K[\psi_\theta^\delta(a, a^*; \boldsymbol{O})]}{\mathbb{P}_K[\delta(N)]} + o_p(K^{-1/2}).$$

By the central limit theorm and the delta method, we conclude that $\sqrt{K}\left\{\widehat{\theta}_V^{\text{eif}_1\text{-ml}}(a, a^*) - \underline{\theta}_V(a, a^*)\right\}$ converges to a normal distribution with mean 0 and variance $\mathbb{E}\left[\left\{\frac{\psi_\theta^\delta(a,a^*;\boldsymbol{O}) - \delta(N)\underline{\theta}_V(a,a^*)}{\mathbb{E}[\delta(N)]}\right\}^2\right] = \mathbb{E}[\{\mathcal{D}_{\theta_V(a,a^*)}(\boldsymbol{O})\}^2]$

□

Supporting information for Remark 2. The following equality is repeatedly used in the proof:

$$\mathbb{E}[M_{\cdot j}|A = a, N] = \mathbb{E}\left\{\mathbb{E}[M_{\cdot j}|A = a, \boldsymbol{C}, N]\Big|N\right\}, \tag{s19}$$

which follows Assumptions 1–2 by observing

$$\begin{aligned} \mathbb{E}\left\{\mathbb{E}[M_{\cdot j}|A = a, \boldsymbol{C}, N]\Big|N\right\} &= \mathbb{E}\left\{\mathbb{E}[M_{\cdot j}(a)|A = a, \boldsymbol{C}, N]\Big|N\right\} \quad \text{(by Assumption 1)} \\ &= \mathbb{E}\left\{\mathbb{E}[M_{\cdot j}(a)|\boldsymbol{C}, N]\Big|N\right\} \quad \text{(by Assumption 2)} \\ &= \mathbb{E}[M_{\cdot j}(a)|N] \quad \text{(by LIE)} \\ &= \mathbb{E}[M_{\cdot j}|A = a, N] \quad \text{(by Assumption 2).} \end{aligned}$$

We first prove $\widehat{\theta}_C^{\text{mf-par}}(a, a^*)$ is consistent under the conditions in Remark 2. Based on the assumptions in Remark 2, we observe that

1. $\widehat{\eta}_{\cdot j}$ is consistently estimated based on a linear regression with the form $\eta_{\cdot j}(A, \boldsymbol{M}, \boldsymbol{C}, N) = \beta_0(A, \boldsymbol{C}, N) + \sum_{l=1}^N \beta_{1,l}(A, N)M_{\cdot l}$.

2. While $\widehat{\kappa}_{\cdot j}$ can be incorrectly specified, the resultant estimator $\widehat{\mathbb{E}}[M_{\cdot j}|A = a^*, N]$ induced from $\widehat{\kappa}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{C}, N)$ is a consistent estimator of $\mathbb{E}[M_{\cdot j}|A = a^*, N]$. This sug-

gests that $\int_{m_{\cdot j}} m_{\cdot j}\widehat{\kappa}_{\cdot j}(a, m_{\cdot j}, \boldsymbol{C}, N)\mathrm{d}m_{\cdot j} = \widehat{\mathbb{E}}[M_{\cdot j}|A = a^*, N]$ is consistent to $\mathbb{E}[M_{\cdot j}|A = a^*, N]$, although it is generally an invalid estimator of $\mathbb{E}[M_{\cdot j}|A = a^*, \boldsymbol{C}, N]$.

Therefore, one can show

$$
\begin{aligned}
\widehat{\theta}_C^{\text{mf-par}}(a, a^*) &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \widehat{\eta}_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N)\widehat{\kappa}^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \left\{\widehat{\beta}_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\widehat{\beta}_{1,l}(a, N)m_{\cdot l}\right\}\widehat{\kappa}^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
&\qquad \text{(by the linear regression expression of } \widehat{\eta}_{\cdot j}) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\widehat{\beta}_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\widehat{\beta}_{1,l}(a, N)\int_{\boldsymbol{m}} m_{\cdot l}\widehat{\kappa}^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right\}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\widehat{\beta}_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\widehat{\beta}_{1,l}(a, N)\int_{m_{\cdot l}} m_{\cdot l}\widehat{\kappa}_{\cdot j}(a^*, m_{\cdot j}, \boldsymbol{C}, N)\mathrm{d}m_{\cdot j}\right\}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\widehat{\beta}_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\widehat{\beta}_{1,l}(a, N)\widehat{\mathbb{E}}[M_{\cdot l}|A = a^*, N]\right\}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\beta_{1,l}(a, N)\mathbb{E}[M_{\cdot l}|A = a^*, N]\right\}\right] + o_p(1),
\end{aligned}
$$

Meanwhile, we have

$$
\begin{aligned}
\theta_C(a, a^*) &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(a, \boldsymbol{m}, \boldsymbol{C}, N)\kappa^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \left\{\beta_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\beta_{1,l}(a, N)m_{\cdot l}\right\}\kappa^c(a^*, \boldsymbol{m}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(a, \boldsymbol{C}, N) + \sum_{l=1}^{N}\beta_{1,l}(a, N)\mathbb{E}[M_{\cdot l}|A = a^*, \boldsymbol{C}, N]\right\}\right].
\end{aligned}
$$

This suggests that

$$
\begin{aligned}
&\widehat{\theta}_C^{\text{mf-par}}(a, a^*) - \theta_C(a, a^*) \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\sum_{l=1}^{N}\beta_{1,l}(a, N)\left\{\mathbb{E}[M_{\cdot l}|A = a^*, N] - \mathbb{E}[M_{\cdot l}|A = a^*, \boldsymbol{C}, N]\right\}\right] + o_p(1) \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\sum_{l=1}^{N}\beta_{1,l}(a, N) \times \underbrace{\mathbb{E}\left\{\mathbb{E}[M_{\cdot l}|A = a^*, N] - \mathbb{E}[M_{\cdot l}|A = a^*, \boldsymbol{C}, N]\middle| N\right\}}_{=0 \text{ due to equation (s19)}}\right] + o_p(1) \\
=&o_p(1).
\end{aligned}
$$

This concludes that $\widehat{\theta}_C^{\text{mf-par}}(a, a^*)$ is consistent to $\theta_C(a, a^*)$. Using a similar strategy, one can demonstrate consistency of $\widehat{\theta}_I^{\text{mf-par}}(a, a^*)$. To prove consistency of $\widehat{\tau}_C^{\text{mf-par}}$, one should note

$$\begin{aligned}
\widehat{\tau}_C^{\text{mf-par}} &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \left\{\widehat{\beta}_0(1, \boldsymbol{C}, N) + \sum_{l=1}^{N}\widehat{\beta}_{1,l}(1, N)m_{\cdot l}\right\}\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\widehat{\beta}_0(1, \boldsymbol{C}, N) + \widehat{\beta}_{1,j}(1, N)\widehat{\mathbb{E}}[M_{\cdot l}|A = 1, N] + \sum_{l\neq j}^{N}\widehat{\beta}_{1,l}(1, N)\widehat{\mathbb{E}}[M_{\cdot l}|A = 0, N]\right\}\right] + o_p(1) \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(1, \boldsymbol{C}, N) + \beta_{1,j}(1, N)\mathbb{E}[M_{\cdot l}|A = 1, N] + \sum_{l\neq j}^{N}\beta_{1,l}(1, N)\mathbb{E}[M_{\cdot l}|A = 0, N]\right\}\right] + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
\tau_C &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}} \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\kappa^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\mathrm{d}\boldsymbol{m}\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(1, \boldsymbol{C}, N) + \beta_{1,j}(1, N)\mathbb{E}[M_{\cdot l}|A = 1, \boldsymbol{C}, N] + \sum_{l\neq j}^{N}\beta_{1,l}(1, N)\mathbb{E}[M_{\cdot l}|A = 0, \boldsymbol{C}, N]\right\}\right],
\end{aligned}$$

which suggests that

$$\begin{aligned}
&\widehat{\tau}_C^{\text{mf-par}} - \tau_C \\
=&\mathbb{E}\Big[\frac{1}{N}\sum_{j=1}^{N}\Big\{\beta_{1,j}(1, N)\left\{\mathbb{E}[M_{\cdot l}|A = 1, N] - \mathbb{E}[M_{\cdot l}|A = 1, \boldsymbol{C}, N]\right\} \\
&\qquad + \sum_{l\neq j}^{N}\beta_{1,l}(1, N)\{\mathbb{E}[M_{\cdot l}|A = 0, N] - \mathbb{E}[M_{\cdot l}|A = 0, \boldsymbol{C}, N]\}\Big\}\Big] + o_p(1) \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_{1,j}(1, N) \times \underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A = 1, N] - \mathbb{E}[M_{\cdot l}|A = 1, \boldsymbol{C}, N]\Big|N\right]}_{=0 \text{ due to equation (s19)}}\right.\right. \\
&\qquad \left.\left. + \sum_{l\neq j}^{N}\beta_{1,l}(1, N)\underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A = 0, N] - \mathbb{E}[M_{\cdot l}|A = 0, \boldsymbol{C}, N]\Big|N\right]}_{=0 \text{ due to equation (s19)}}\right\}\right] + o_p(1) \\
=&o_p(1).
\end{aligned}$$

This confirms the consistency of $\widehat{\tau}_C^{\text{mf-par}}$. Based on a similar strategy, we can show consistency of $\widehat{\tau}_I^{\text{mf-par}}$.

Now, we consider consistency of $\widehat{\zeta}^{\text{eif}_1\text{-par}}$ for $\zeta \in \{\theta_C(a, a^*), \theta_I(a, a^*), \tau_C, \tau_I\}$. The double robustness property in Theorem 3 already suggests that $\widehat{\theta}_C^{\text{eif}_1\text{-par}}(a, a^*)$ and $\widehat{\theta}_I^{\text{eif}_1\text{-par}}(a, a^*)$ are consistent if $\eta_{\cdot j}$ is correctly specified, regardless of whether $\kappa_{\cdot j}$ is correct or not. We therefore only need to prove consistency of $\widehat{\tau}_C^{\text{eif}_1\text{-par}}$ and $\widehat{\tau}_I^{\text{eif}_1\text{-par}}$. Next, we demonstrate consistency of $\widehat{\tau}_C^{\text{eif}_1\text{-par}}$, where consistency of $\widehat{\tau}_I^{\text{eif}_1\text{-par}}$ can be similarly obtained. By Lemma

1, one can show that

$$
\begin{aligned}
&\widehat{\tau}_C^{\text{eif}_1\text{-par}} - \tau_C \\
=&\mathbb{E}[\widehat{\psi}_\tau(\boldsymbol{O}) - \psi_\tau(\boldsymbol{O})] + o_p(1) \\
=&\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \frac{\widehat{\kappa}_{\cdot j}(1, m_j, \boldsymbol{C}, N)\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)}{\widehat{\kappa}^c(1, \boldsymbol{m}, \boldsymbol{C}, N)} \{\widehat{\kappa}^c(1, \boldsymbol{m}, \boldsymbol{C}, N) - \kappa^c(1, \boldsymbol{m}, \boldsymbol{C}, N)\}\{\widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\}\,\mathrm{d}\boldsymbol{m}\Bigg] \\
&- \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\{\widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\}\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\}\,\mathrm{d}\boldsymbol{m}\right] \\
&- \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\{\widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N) - \eta_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\}\left\{\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N) - \kappa^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\right\}\mathrm{d}\boldsymbol{m}\right] \\
&- \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\}\,\mathrm{d}\boldsymbol{m}\right] \\
&+ \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \widehat{\eta}_{\cdot j}(1, \boldsymbol{m}, \boldsymbol{C}, N)\kappa^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\}\,\mathrm{d}\boldsymbol{m}\right] + o_p(1) \\
=&\Delta_1 - \Delta_2 - \Delta_3 - \Delta_4 + \Delta_5 + o_p(1).
\end{aligned}
$$

Here, $\Delta_1 = \Delta_2 = \Delta_3 = o_p(1)$ because we assume $\eta_{\cdot j}$ is correctly specified. To analyze Δ_4, we can show

$$
\begin{aligned}
\Delta_4 =&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \int_{\boldsymbol{m}} \left\{\widehat{\beta}_0(1, \boldsymbol{C}, N) + \sum_{l=1}^N \widehat{\beta}_{1,l}(1, N)m_{\cdot l}\right\}\widehat{\kappa}^c_{\cdot(-j)}(0, \boldsymbol{m}_{\cdot(-j)}, \boldsymbol{C}, N)\{\widehat{\kappa}_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N) - \kappa_{\cdot j}(1, m_{\cdot j}, \boldsymbol{C}, N)\}\,\mathrm{d}\boldsymbol{m}\right] \\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \left\{\widehat{\beta}_0(1, \boldsymbol{C}, N) + \widehat{\beta}_{1,j}(1, N)\widehat{\mathbb{E}}[M_{\cdot j}|A=1, N] + \sum_{l\neq j}\widehat{\beta}_{1,l}(1, N)\widehat{\mathbb{E}}[M_{\cdot j}|A=0, N]\right\}\right] \\
&- \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \left\{\widehat{\beta}_0(1, \boldsymbol{C}, N) + \widehat{\beta}_{1,j}(1, N)\mathbb{E}[M_{\cdot j}|A=1, \boldsymbol{C}, N] + \sum_{l\neq j}\widehat{\beta}_{1,l}(1, N)\mathbb{E}[M_{\cdot j}|A=0, \boldsymbol{C}, N]\right\}\right] \\
=&\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^N \Bigg\{\beta_{1,j}(1, N) \times \underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A=1, N] - \mathbb{E}[M_{\cdot l}|A=1, \boldsymbol{C}, N]\middle|N\right]}_{=0 \text{ due to equation (s19)}} \\
&\qquad + \sum_{l\neq j}^N \beta_{1,l}(1, N)\underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A=0, N] - \mathbb{E}[M_{\cdot l}|A=0, \boldsymbol{C}, N]\middle|N\right]}_{=0 \text{ due to equation (s19)}}\Bigg\}\Bigg] + o_p(1) \\
=&o_p(1).
\end{aligned}
$$

Similarly, we can show $\Delta_5 = o_p(1)$, which concludes that $\widehat{\tau}_C^{\text{eif}_1\text{-par}} = \tau_C + o_p(1)$.

At the end, we prove consistency of $\widehat{\zeta}^{\text{eif}_2\text{-par}}$ for $\zeta \in \{\theta_C(a, a^*), \theta_I(a, a^*), \tau_C, \tau_I\}$. Notice that consistency of $\widehat{\theta}_C^{\text{eif}_2\text{-par}}(a, a^*)$ and $\widehat{\theta}_I^{\text{eif}_2\text{-par}}(a, a^*)$ automatically holds because they do not involve the working model of $\kappa_{\cdot j}$ and are consistent if $\eta_{\cdot j}$ is correctly specified (Theorem 3). For consistency of $\widehat{\tau}_C^{\text{eif}_2\text{-par}}$, one can show

$$
\begin{aligned}
&\widehat{\tau}_C^{\text{eif}_2\text{-par}} - \tau_C \\
=&\mathbb{E}[\widehat{\psi}_\tau(\boldsymbol{O}) - \psi_\tau(\boldsymbol{O})] + o_p(1) \\
&\text{(by Lemma 2, we then have)}
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{s(1,\boldsymbol{M},\boldsymbol{C},N)}{f(A=0)}\left\{\frac{\widehat{\kappa}_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\widehat{\kappa}^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{\widehat{s}(0,\boldsymbol{M},\boldsymbol{C},N)}{\widehat{s}(1,\boldsymbol{M},\boldsymbol{C},N)}-\frac{\kappa_{\cdot j}(1,M_{\cdot j},\boldsymbol{C},n)}{\kappa^{\star}_{\cdot j}(0,\boldsymbol{M},\boldsymbol{C},N)}\frac{s(0,\boldsymbol{M},\boldsymbol{C},N)}{s(1,\boldsymbol{M},\boldsymbol{C},N)}\right\}\{\eta_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)-\widehat{\eta}_{\cdot j}(1,\boldsymbol{M},\boldsymbol{C},N)\}\right]\\
&+\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\{\widehat{\eta}_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)-\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\}\,\mathrm{d}\boldsymbol{m}\right]\\
&+\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\eta^{\dagger}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}m_{\cdot j}\right]\\
&-\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{\eta}^{\dagger}_{\cdot j}(1,0,m_{\cdot j},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}m_{\cdot j}\right]+o_p(1)\\
=&\Delta_1+\Delta_2+\Delta_3-\Delta_4+o_p(1),
\end{aligned}
$$

where $\Delta_1=\Delta_2=o_p(1)$ due to the assumption that $\eta_{\cdot j}$ is correctly specified. Because $\eta^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)=\int_{\boldsymbol{m}_{\cdot(-j)}}\eta_{\cdot j}(a,\boldsymbol{m},\boldsymbol{c},n)\kappa_{\cdot(-j)}(a^*,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{c},n)\mathrm{d}\boldsymbol{m}_{\cdot(-j)}$, we obtain

$$
\begin{aligned}
\Delta_3=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{\boldsymbol{m}}\eta_{\cdot j}(1,\boldsymbol{m},\boldsymbol{C},N)\kappa_{\cdot(-j)}(0,\boldsymbol{m}_{\cdot(-j)},\boldsymbol{C},N)\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}\boldsymbol{m}\right]\\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(1,\boldsymbol{C},N)+\beta_{1,j}(1,N)\mathbb{E}[M_{\cdot l}|A=1,N]+\sum_{l\neq j}^{N}\beta_{1,l}(1,N)\mathbb{E}[M_{\cdot l}|A=0,N]\right\}\right]\\
&-\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\{\beta_0(1,\boldsymbol{C},N)+\beta_{1,j}(1,N)\mathbb{E}[M_{\cdot l}|A=1,\boldsymbol{C},N]+\sum_{l\neq j}^{N}\beta_{1,l}(1,N)\mathbb{E}[M_{\cdot l}|A=0,\boldsymbol{C},N]\right\}\right]+o_p(1)\\
=&\mathbb{E}\Bigg[\frac{1}{N}\sum_{j=1}^{N}\Bigg\{\beta_{1,j}(1,N)\times\underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A=1,N]-\mathbb{E}[M_{\cdot l}|A=1,\boldsymbol{C},N]\Big|N\right]}_{=0\text{ due to equation (s19)}}\\
&\qquad+\sum_{l\neq j}^{N}\beta_{1,l}(1,N)\underbrace{\mathbb{E}\left[\mathbb{E}[M_{\cdot l}|A=0,N]-\mathbb{E}[M_{\cdot l}|A=0,\boldsymbol{C},N]\Big|N\right]}_{=0\text{ due to equation (s19)}}\Bigg\}\Bigg]+o_p(1)\\
=&o_p(1).
\end{aligned}
$$

To analyze Δ_4, recall that we fit a linear regression on $\eta^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$ with no interaction terms between $M_{\cdot j}$ and $\boldsymbol{C}$. Without loss of generality, we consider that $\eta^{\dagger}_{\cdot j}(a,a^*,m_{\cdot j},\boldsymbol{c},n)$ can be fitted by

$$
\eta^{\dagger}_{\cdot j}(a,a^*,M_{\cdot j},\boldsymbol{C},N)=\psi_0(a,a^*,N,\boldsymbol{C})+\psi_1(a,a^*,N)M_{\cdot j}, \tag{s20}
$$

where $\psi_0(a,a^*,N,\boldsymbol{C})$ and $\psi_1(a,a^*,N)$ are coefficients of $\{a,a^*,N,\boldsymbol{C}\}$ and $\{a,a^*,N\}$, respectively. Note that we do not assume (s20) to be correctly specified. Due to potential misspecification, $\widehat{\psi}_0(a,a^*,N,\boldsymbol{C})$ and $\widehat{\psi}_1(a,a^*,N)$ may converge to $\widetilde{\psi}_0(a,a^*,N,\boldsymbol{C})$ and $\widetilde{\psi}_1(a,a^*,N)$ that are different from their true values. Observing this, we can write out Δ_4 as

$$
\begin{aligned}
\Delta_4=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\int_{m_{\cdot j}}\widehat{\psi}_1(a,a^*,N)m_{\cdot j}\times\{\widehat{\kappa}_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)-\kappa_{\cdot j}(1,m_{\cdot j},\boldsymbol{C},N)\}\,\mathrm{d}m_{\cdot j}\right]\\
=&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\widehat{\psi}_1(a,a^*,N)\times\left\{\widehat{\mathbb{E}}[M_{\cdot j}|A=1,N]-\mathbb{E}[M_{\cdot j}|A=1,\boldsymbol{C},N]\right\}\right]
\end{aligned}
$$

(because $\widehat{\psi}_1(a, a^*, N)$ converges to $\widetilde{\psi}_1(a, a^*, N)$ and the GLM of $\mathbb{E}[M_{\cdot j}|A=1,N]$ is correctly specified, we then have)

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\widetilde{\psi}_1(a,a^*,N)\times\{\mathbb{E}[M_{\cdot j}|A=1,N]-\mathbb{E}[M_{\cdot j}|A=1,\boldsymbol{C},N]\}\right]+o_p(1)$$

$$=\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\widetilde{\psi}_1(a,a^*,N)\times\underbrace{\mathbb{E}\left\{\mathbb{E}[M_{\cdot j}|A=1,N]-\mathbb{E}[M_{\cdot j}|A=1,\boldsymbol{C},N]\middle|N\right\}}_{=0\text{ due to equation (s19)}}\right]+o_p(1)$$

$$=o_p(1).$$

This concludes the consistency of $\widehat{\tau}_C^{\text{eif}_2\text{-par}}$. Based on a similar strategy, we can show consistency of $\widehat{\tau}_I^{\text{eif}_2\text{-par}}$. □

Proof of Corollary 1. Resume our notation in the proof of Theorem 2. Define $\zeta_1(t)$ and $\zeta_2(t)$ as the value of ζ_1 and ζ_2 evaluated under the parametric submodel $f_t(\boldsymbol{O})$. Because $\mathcal{D}_{\zeta_1}(\boldsymbol{O})$ and $\mathcal{D}_{\zeta_2}(\boldsymbol{O})$ are the EIFs of ζ_1 and ζ_2, we have that

$$\mathbb{E}\left[\mathcal{D}_{\zeta_1}(\boldsymbol{O})S(\boldsymbol{O})\right]=\nabla_{t=0}\zeta_1(t)\text{ and }\mathbb{E}\left[\mathcal{D}_{\zeta_2}(\boldsymbol{O})S(\boldsymbol{O})\right]=\nabla_{t=0}\zeta_2(t). \tag{s21}$$

Also, because $\Delta=g(\zeta_1,\zeta_2)$, the nonparametric identification formula of Δ under the parametric submodel is $\Delta(t)=g(\zeta_1(t),\zeta_2(t))$. Notice that

$$\begin{aligned}\nabla_{t=0}\Delta(t)&=\dot{g}_1(\zeta_1,\zeta_2)\nabla_{t=0}\zeta_1(t)+\dot{g}_2(\zeta_1,\zeta_2)\nabla_{t=0}\zeta_2(t)\\&=\mathbb{E}\left[\{\dot{g}_1(\zeta_1,\zeta_2)\mathcal{D}_{\zeta_1}(\boldsymbol{O})+\dot{g}_2(\zeta_1,\zeta_2)\mathcal{D}_{\zeta_2}(\boldsymbol{O})\}\,S(\boldsymbol{O})\right]\\&=\mathbb{E}\left[\mathcal{D}_{\Delta}(\boldsymbol{O})S(\boldsymbol{O})\right],\end{aligned}$$

where the first equality follows the chain rule and the second equality follows (s21). This concludes that $\mathcal{D}_\Delta(\boldsymbol{O})$ is the EIF of Δ. □

Proof of Corollaries 2 and 3. The conclusions in Corollaries 2 and 3 can be easily verified by using the results in Theorem 3 and 4, respectively. □

C Supplementary Material Figures

The Supplementary Material Figures are summarized as follows:

1. Figure 1: Simulation results of NIE_I in the scenario with $K=100$ clusters.
2. Figure 2: Simulation results of SME_I in the scenario with $K=100$ clusters.
3. Figure 3: Simulation results of NIE_C in the scenario with $K=50$ clusters.
4. Figure 4: Simulation results of SME_C in the scenario with $K=50$ clusters.
5. Figure 5: Simulation results of NIE_I in the scenario with $K=50$ clusters.
6. Figure 6: Simulation results of SME_I in the scenario with $K=50$ clusters.

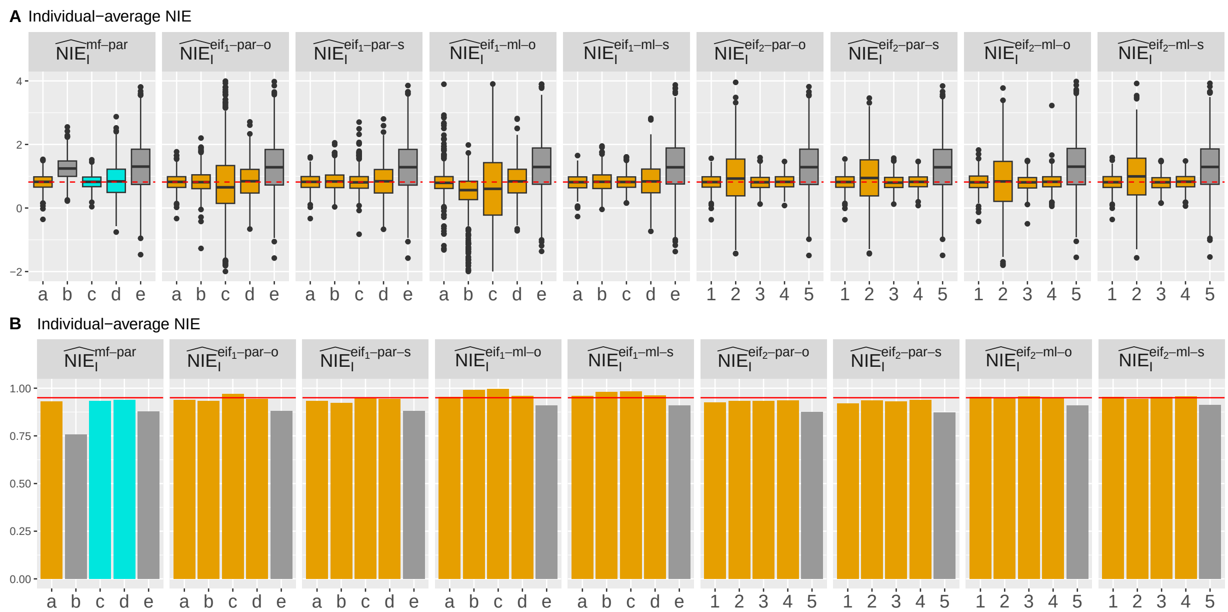

Figure 1: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of NIE_I, in the scenario with $K = 100$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ incorrect.

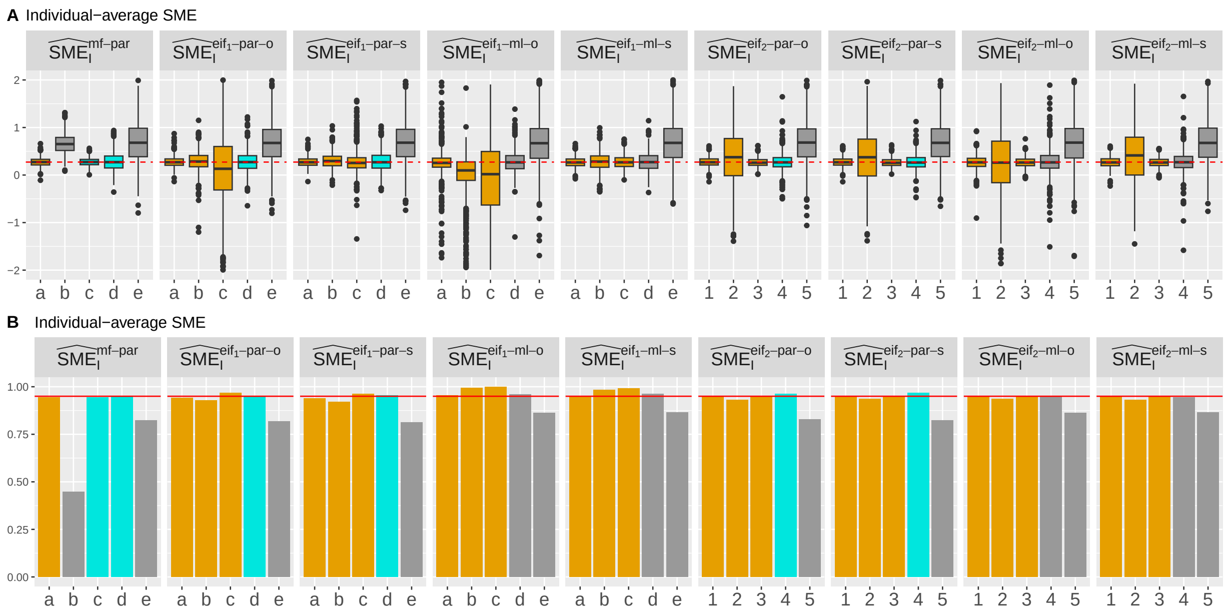

Figure 2: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of SME_I, in the scenario with $K = 100$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ incorrect.

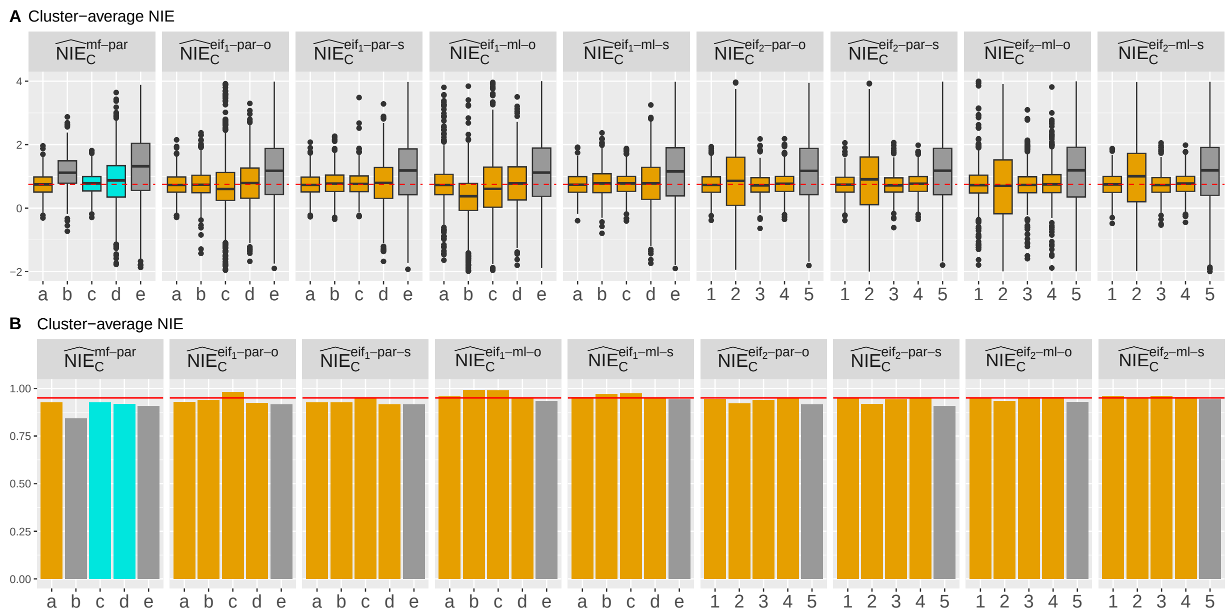

Figure 3: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of NIE_C, in the scenario with $K = 50$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ incorrect.

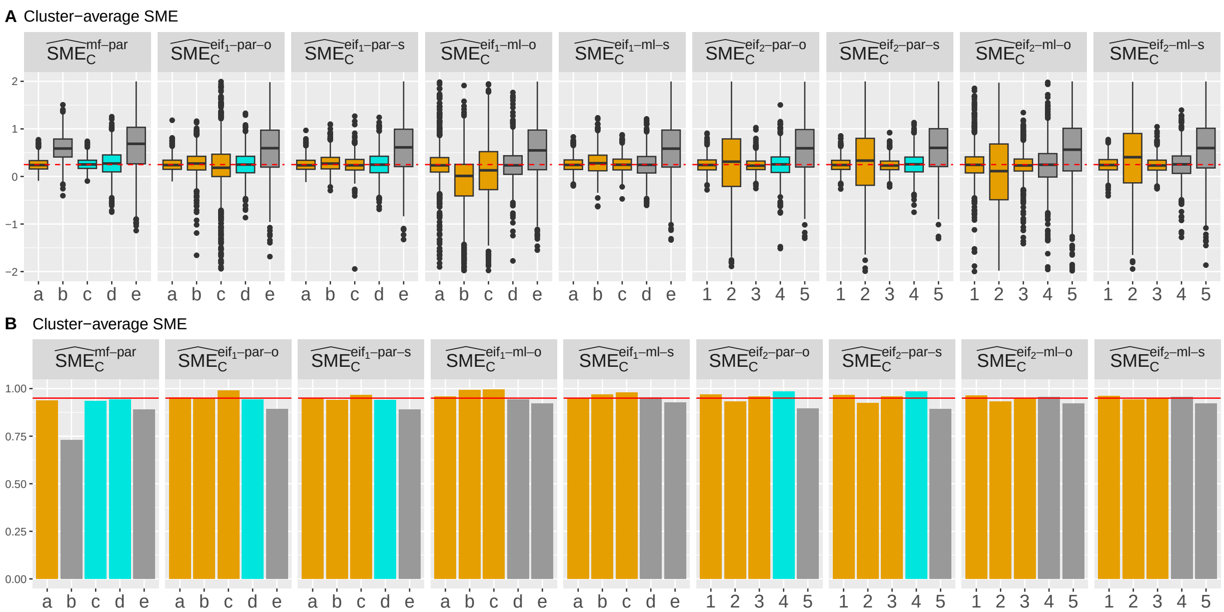

Figure 4: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of SME_C, in the scenario with $K = 50$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ incorrect.

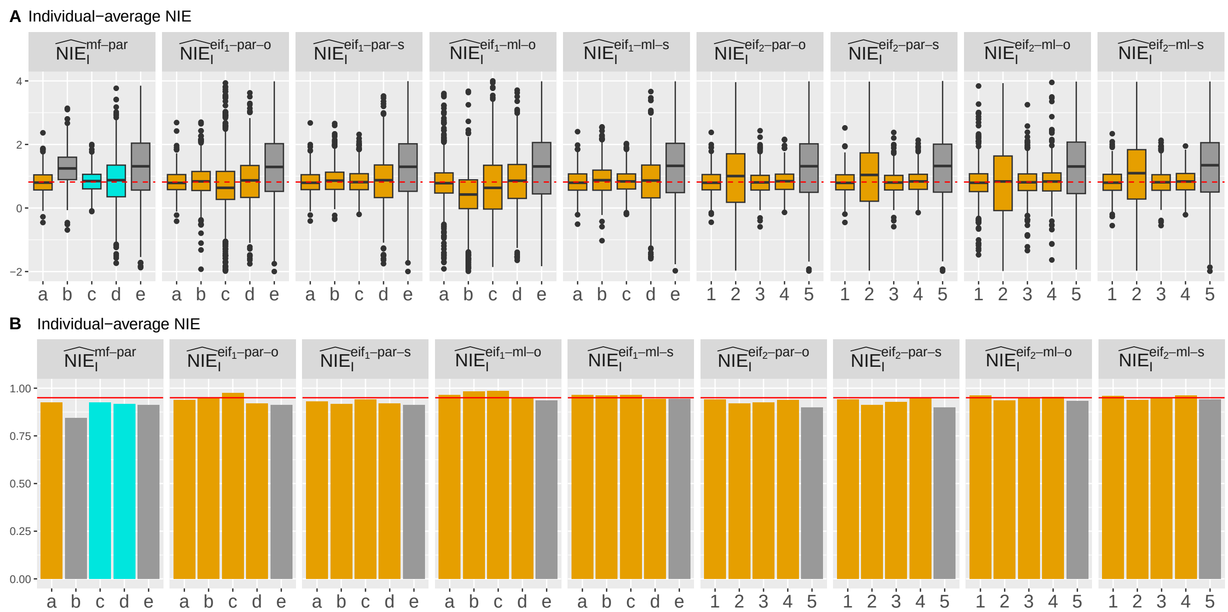

Figure 5: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of NIE_I, in the scenario with $K = 50$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^\star_{\cdot j}, \eta^\dagger_{\cdot j}, \kappa_{\cdot j}, \kappa^\star_{\cdot j}, s\}$ incorrect.

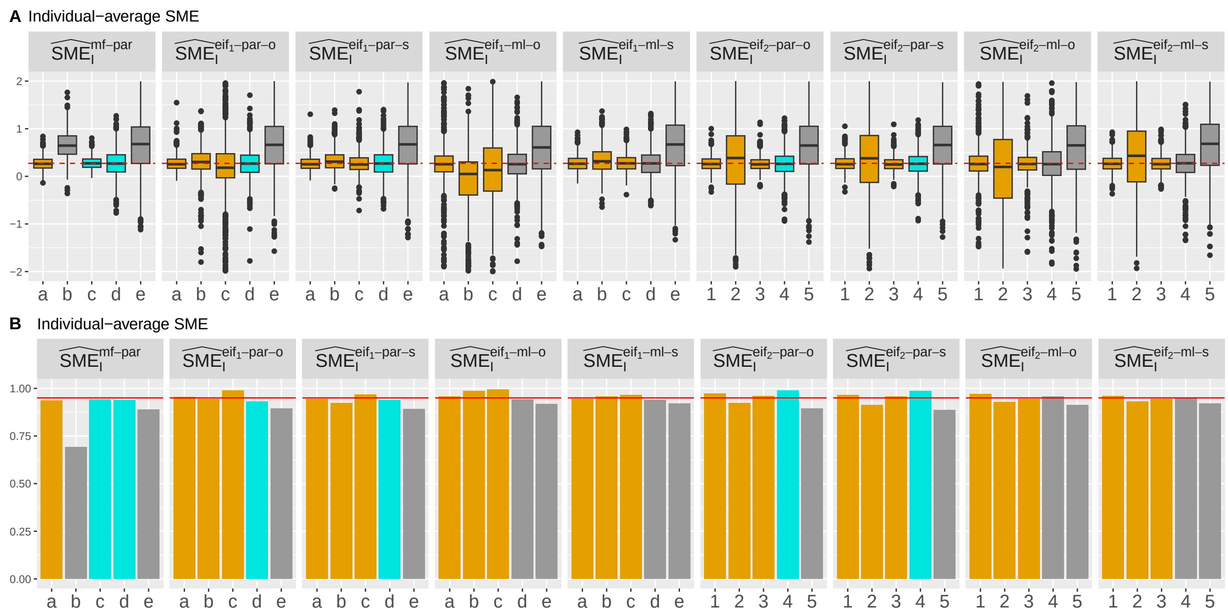

Figure 6: Sampling distributions (Panel A) and 95% Confidence Interval Coverage Probability (Panel B) among estimators of SME_I, in the scenario with $K = 50$ clusters. For each scenario, the box/bar filled with orange color indicates that the estimator is consistent based on theory, and the box/bar filled with blue color indicates that the estimator is consistent based on Remark 2. For methods require specifying $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$, we consider Scenarios (a) all $\{\kappa_{\cdot j}, \eta_{\cdot j}, \mathcal{C}\}$ correct, (b) $\{\kappa_{\cdot j}, \mathcal{C}\}$ correct, (c) $\{\kappa_{\cdot j}, \eta_{\cdot j}\}$ correct, (d) $\{\eta_{\cdot j}, \mathcal{C}\}$ correct, and (e) $\mathcal{C}$ correct. For methods require specifying $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$, we consider Scenarios (1) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ correct, (2) only $\{s, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (3) only $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}\}$ correct, (4) only $\{s, \eta_{\cdot j}\}$ correct, (5) all $\{\eta_{\cdot j}, \eta^{\star}_{\cdot j}, \eta^{\dagger}_{\cdot j}, \kappa_{\cdot j}, \kappa^{\star}_{\cdot j}, s\}$ incorrect.