

Flexible Nonparametric Inference for Causal Effects under the Front-Door Model

Anna Guo, David Benkeser, Razieh Nabi

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA

anna.guo@emory.edu, benkeser@emory.edu, razieh.nabi@emory.edu

Abstract

Evaluating causal treatment effects in observational studies requires addressing confounding. While the *back-door* criterion enables identification through adjustment for observed covariates, it fails in the presence of unmeasured confounding. The *front-door* criterion offers an alternative by leveraging variables that fully mediate the treatment effect and are unaffected by unmeasured confounders of the treatment-outcome pair. We develop novel one-step and targeted minimum loss-based estimators for both the *average treatment effect* and the *average treatment effect on the treated* under front-door assumptions. Our estimators are built on multiple parameterizations of the observed data distribution, including approaches that avoid modeling the mediator density entirely, and are compatible with flexible, machine learning-based nuisance estimation. We establish conditions for root- n consistency and asymptotic linearity by deriving second-order remainder bounds. We also develop flexible tests for assessing identification assumptions, including a doubly robust testing procedure, within a semiparametric extension of the front-door model that encodes *generalized (Verma) independence* constraints. We further show how these constraints can be leveraged to improve the efficiency of causal effect estimators. Simulation studies confirm favorable finite-sample performance, and real-data applications in education and emergency medicine illustrate the practical utility of our methods. An accompanying R package, `fdcausal`, implements all proposed procedures.

Keywords: Unmeasured confounders, Double-debiased machine learning, Model evaluation

1 Introduction

Two key causal parameters are the average treatment effect (ATE), which captures the population-level causal effect, and the average treatment effect on the treated (ATT), which captures the effect within the subpopulation that naturally receives treatment. When all confounders are observed, identification of these effects is often achieved using the *back-door* criterion, which involves adjusting for a set of covariates that block all non-causal paths between the treatment and outcome [Pearl, 2009]. Under this criterion, the ATE is identified via the g-formula [Robins, 1986, Hahn, 1998] and/or inverse probability of treatment weighting (IPTW) [Hirano et al., 2003]. A rich literature exists for estimating these functionals using plug-in, IPTW, augmented IPTW, and targeted minimum loss-based estimators (TMLEs) [Bickel et al., 1993, van der Vaart, 2000, Tsiatis, 2007, Robins et al., 1994, van der Laan et al., 2011, Chernozhukov et al., 2017].

Identifying a sufficient back-door adjustment set is not always feasible in practice due to unmeasured confounding. In such settings, a variety of alternative strategies have been proposed, including instrumental variable methods [Balke and Pearl, 1994], sensitivity analyses [Robins et al., 2000, Scharfstein et al., 2021], and bounds analysis [Manski, 1990]. Other approaches include those based on causal graphical models that enable reasoning about identification using independence constraints between counterfactual and observed variables [Tian and Pearl, 2002, Richardson and Robins, 2013]. These models underlie *sound* and *complete* algorithms for identifying causal parameters from observed data [Shpitser and Pearl, 2006, Huang and Valtorta, 2006, Bhattacharya et al., 2022, Richardson et al., 2023].

The *front-door* criterion is an identification strategy that enables inference even in the presence of unmeasured confounding [Pearl, 1995]. This criterion requires the existence of one or more mediators that satisfy two key conditions: (i) no unmeasured confounding between the treatment and mediators nor between mediators and outcome, and (ii) the effect of treatment on the outcome is fully mediated through the mediators. When these conditions hold, average causal effects are identifiable from observed data. In settings where the full mediation assumption (ii) is violated, Fulcher et al. [2019] proposed the *population intervention indirect effect*, which relaxes this assumption by introducing an additional *cross-world counterfactual* independence. Although the estimand differs, the underlying identification strategy remains closely related. Empirical

studies suggest the front-door criterion can yield reliable estimates in real-world settings where unmeasured confounding is expected [Glynn and Kashin, 2018, Bellemare et al., 2019, Fulcher et al., 2019, Bhattacharya and Nabi, 2022, Piccininni et al., 2023, Wen et al., 2024].

A nonparametric efficient estimator of the front-door functional was proposed by Fulcher et al. [2019], who developed a *one-step estimator* based on parametric working models for three key nuisance components: the conditional mean outcome, the conditional density of the mediator, and the conditional probability of treatment. This estimator enjoys the property of *double robustness* and marked an important contribution to front-door estimation. However, several critical gaps in its applicability remain. First, its reliance on parametric modeling restricts applicability in settings that demand flexible, data-adaptive nuisance estimation. Second, the approach is functionally restricted to settings with a single mediator, as it requires estimation of the mediator density. Yet in practice, multiple mediators often arise—whether to satisfy identification assumptions under full mediation or to capture complex indirect pathways under partial mediation—making density estimation impractical. Third, the one-step estimator can produce estimates outside of the natural parameter space, which is particularly concerning for binary or bounded continuous outcomes. Recent work by Wen et al. [2024] addresses some of these issues by introducing TMLE-based estimators for a related target parameter, using a reparameterization of the efficient influence function that avoids direct modeling of the mediator density. Their approach improves practical feasibility, especially in settings with continuous mediators. However, their estimand differs from the standard ATE front-door functional studied here, and questions remain as to whether and how to incorporate flexible nuisance estimation into that framework.

The front-door criterion enables identification of causal effects under unmeasured treatment-outcome confounding assuming the absence of unmeasured confounding between treatment and mediator(s), between mediator(s) and outcome, and the absence of a direct effect of treatment on outcome. However, these assumptions are themselves untestable in a nonparametrically saturated model. Bhattacharya and Nabi [2022] described the use of an auxiliary *anchor* variable, a baseline covariate associated with treatment (and possibly mediator) but not a direct cause of the outcome, to assess the front-door assumptions. The presence of such an anchor induces a testable *Verma constraint*—a type of generalized independence relation in the observed data distribution

[Verma and Pearl, 1990]—that encodes the absence of a direct effect of the anchor variable on the outcome. Parametric tests for this constraint have been proposed, but they rely on strong modeling assumptions and are limited in flexibility.

This work extends the foundational contributions of Fulcher et al. [2019], Wen et al. [2024], and Bhattacharya and Nabi [2022] in several respects. First, we propose a suite of robust and efficient estimators for the ATE front-door functional based on three different parameterizations of the observed data distribution that enable scalable and robust inference regarding the front door functional in the presence of multivariate mediators of mixed types (Section 3). Second, we develop efficient estimators of the ATT under the front-door model, which have previously not been described in literature (Section 4). Third, we derive and analyze second-order remainder terms for all proposed ATE and ATT estimators and establish conditions under which root- n consistency and asymptotic linearity hold when nuisance functions are estimated using flexible, data-adaptive methods (Section 5). Characterizing these remainder terms lays the foundation for additional work in increasing the robustness of confidence interval and hypothesis test construction [Van der Laan, 2014, Benkeser et al., 2017]. Fourth, we evaluate the validity of the front-door model with an anchor variable by developing flexible tests based on weighted risk minimization, along with a novel doubly robust testing procedure (Sections 6.1 and 6.2). We further show how the Verma constraint can be exploited to improve efficiency of causal effect estimators (Section 6.3). Finally, we demonstrate the practical utility of our methods through extensive simulation studies (Section 7) and two diverse real-world applications: one analyzing the effect of early academic performance on later income and another evaluating the impact of mobile stroke unit deployment on clinical outcomes in emergency medicine (Section 8). An R package, `fdcausal` implementing all the proposed methods is publicly available on Github at [annaguo-bios/fdcausal](https://github.com/annaguo-bios/fdcausal).

2 Causal front-door model

Let A denote the observed treatment and Y denote the observed outcome of interest. We assume the treatment is binary, with $A = 1$ representing the treatment arm and $A = 0$ representing the control arm. We use Y^a to denote the potential outcome if the treatment variable was assigned the value $a \in \{0, 1\}$ [Neyman, 1923, Rubin, 1974]. We write P for distributions and p for

densities, assuming continuous variables admit Lebesgue densities (though this is not required). The ATE and ATT are defined as $\text{ATE} := \mathbb{E}(Y^1 - Y^0)$ and $\text{ATT} := \mathbb{E}(Y^1 - Y^0 | A = 1)$, where $\mathbb{E}(Y^a) = \int y p(Y^a = y) dy$ and $\mathbb{E}(Y^a | A = 1) = \int y p(Y^a = y | A = 1) dy$.

Common identification approaches assume: (i) *consistency* which states that $Y = AY^1 + (1 - A)Y^0$; (ii) *conditional ignorability* which assumes the existence of a set of observed pre-treatment covariates X such that $Y^a \perp\!\!\!\perp A | X$, for $a \in \{0, 1\}$; and (iii) *positivity* which ensures that $p(A = a | x) > 0$ for $a \in \{0, 1\}$ and all x in the support of X . Under assumptions (i)-(iii), the ATE and ATT are both identified via the *back-door adjustment formulae* $\mathbb{E}(\mathbb{E}(Y | A = 1, X) - \mathbb{E}(Y | A = 0, X))$ and $\mathbb{E}(\mathbb{E}(Y | A = 1, X) - \mathbb{E}(Y | A = 0, X)) | A = 1$, respectively. We note that ATT identification requires a weaker form of positivity: $p(A = 0 | x) > 0$ for all x such that $p(A = 1 | x) > 0$. This causal model corresponds to the DAG in Fig. 1(a) (without $A \leftarrow U \rightarrow Y$ edges).

Various methods have been developed to infer the back-door adjustment formulae from observed data, including propensity score matching [Rosenbaum and Rubin, 1983], g-computation [Robins, 1986], (stabilized) IPTW [Hernán and Robins, 2006], augmented IPTW [Robins et al., 1994], and TMLE [van der Laan and Rubin, 2006]. However, in the presence of unmeasured confounders (U in Fig. 1(a)), the ATE and ATT are no longer identifiable, and any inference based on the back-door adjustment formulae are likely to be biased.

As an alternative to the back-door, Pearl proposed the front-door model [Pearl, 1995], which enables causal identification even in the presence of unmeasured confounders. The core idea is to identify a vector of mediators M that intercept all directed paths from A to Y and that share no unmeasured confounders with either the treatment or the outcome. These conditions correspond to the absence of dashed gray edges in Fig. 1(b), where U_{AM} and U_{MY} encode unmeasured confounding sources between the treatment-mediator and mediator-outcome pairs, respectively. The focus of our work is a generalized version of front-door model that allows for the existence of observed common causes X between treatment, mediator, and outcome (Fig. 1(c)).

2.1 Identification of the ATE and ATT

The identification assumptions for ATE in the front-door model based on observations of $O = (X, A, M, Y) \sim P$ are as follows: (i) *consistency* which states $M^a = M$ when $A = a$ and

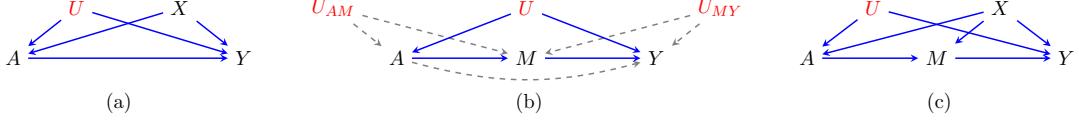


Figure 1: (a) Example of a DAG with measured confounders X and unmeasured confounders U ; (b) The front-door DAG with unmeasured confounders U between A and Y (dashed edges indicate assumptions); (c) The front-door DAG with the inclusion of measured confounders X .

$Y^m = Y$ when $M = m$; (ii) *conditional ignorability* which assumes the absence of unmeasured confounders between the treatment-mediator and mediator-outcome pairs, i.e., $M^a \perp A | X$ and $Y^m \perp M | A, X$; (iii) *no direct effect* which assumes that M intercepts all directed paths from A to Y , i.e., $Y^{a,m} = Y^m$ for $a \in \{0, 1\}$ and all m in the support of M ; and (iv) *positivity* which ensures that $p(A = a | X = x)$ and $p(M = m | A = a, X = x)$ are positive for all (x, a, m) in the support of (X, A, M) . We denote by \mathcal{M} our model for the observed data distribution P , which is nonparametric up to the positivity conditions in (iv).

Given that identification arguments and estimation techniques for $\mathbb{E}(Y^1)$ and $\mathbb{E}(Y^0)$ are similar, we explicitly consider $\mathbb{E}(Y^{a_0})$, $a_0 \in \{0, 1\}$ to be the parameter of interest when studying the ATE. Under assumptions (i)-(iv), $\mathbb{E}(Y^{a_0})$ is identified by $\psi_{a_0}(P)$ [Pearl, 1995], where

$$\psi_{a_0}(P) = \iiint \sum_{a=0}^1 y p(y | m, a, x) p(a | x) p(m | A = a_0, x) p(x) dy dm dx. \quad (1)$$

Under the same assumptions, the ATT can also be expressed as a functional of P . To enable a formulation that naturally extends to the *average treatment effect among controls* (ATC), defined as $\mathbb{E}(Y^1 - Y^0 | A = 0)$, we consider the general counterfactual quantity $\mathbb{E}(Y^{a_0} | A = 1 - a_0)$, for $a_0 \in \{0, 1\}$. Since $\mathbb{E}(Y^{a_0} | A = a_0)$ is identified by consistency as $\mathbb{E}(Y | A = a_0)$ and can be directly estimated via the subpopulation sample mean, we focus on the nontrivial term $\mathbb{E}(Y^{a_0} | A = 1 - a_0)$, which is identified under the front-door model by the functional:

$$\beta_{a_0}(P) = \iiint y p(y | m, A = 1 - a_0, x) p(m | A = a_0, x) p(x | A = 1 - a_0) dy dm dx. \quad (2)$$

We note that above identification requires a weaker form of positivity: $p(M = m | A = a, X = x) > 0$ for $a \in \{0, 1\}$, all m in the support of M , and all x such that $p(X = x | A = 1 - a_0) > 0$.

We adopt the terminology of ATE and ATT front-door functionals to refer to ψ_{a_0} and β_{a_0} , respectively, with the understanding that these represent counterfactual means rather than effect

contrasts. Under these formulations, the ATE, ATT, and ATC are identified as $\psi_1(P) - \psi_0(P)$, $\mathbb{E}(Y|A=1) - \beta_0(P)$, and $\beta_1(P) - \mathbb{E}(Y|A=0)$, respectively (see Appendix B.1 for proof).

Alternative interpretations of the front-door functionals: The ATE front-door functional in (1) admits multiple, closely related causal interpretations beyond the usual full-mediation setting. In particular, it coincides with the *population intervention indirect effect* (PIIE) of Fulcher et al. [2019], defined as $\mathbb{E}(Y - Y^{A,M^{a_0}})$ and identifiable under a cross-world independence assumption rather than the no-direct-effect assumption. We note that the ATT front-door functional in (2) can also recover subgroup-specific PIIEs (among treated or controls). A further framing by Wen et al. [2024] regards the same functional as the *average causal effect of an intervenable treatment component* A_M , namely $\mathbb{E}(Y^{a_M=1} - Y^{a_M=0})$, which is identified by the front-door formula even when A itself is not manipulable. Thus, our estimators for both ATE and ATT continue to estimate meaningful indirect effects whenever the full mediation assumption is relaxed or when one targets modifiable treatment components. For detailed discussion, see Appendix B.2.

Our primary objective is to develop estimators for the front-door functionals in (1) and (2), using n i.i.d. observations of $O = (X, A, M, Y)$. We begin by reviewing existing estimation strategies for the ATE front-door functional and highlighting their limitations. In contrast, estimation results for the ATT front-door functional have received little to no prior attention.

2.2 Prior estimation for the ATE front-door functional

Let $Q = (\mu, \pi, f_M, p_X)$ denote the collection of key *nuisance parameters*, where $\mu(m, a, x) = \mathbb{E}(Y | M = m, A = a, X = x)$, $\pi(a|x) = P(A = a | X = x)$, $f_M(m | a_0, x) = p(M = m | A = a_0, X = x)$, and $p_X(x) = p(X = x)$. Then $\psi_{a_0}(P)$ and $\beta_{a_0}(P)$ can equivalently be written as $\psi_{a_0}(Q)$ and $\beta_{a_0}(Q)$ for a fixed choice of $a_0 \in \{0, 1\}$. To simplify notation, we suppress the subscript a_0 going forward. It is also useful for our later developments to introduce the following quantities: $\xi(M, X) := \sum_{a=0}^1 \mu(M, a, X) \pi(a | X)$, $\eta(A, X) := \int \mu(m, A, X) f_M(m | a_0, X) dm$, and $\theta(X) := \int \xi(m, X) f_M(m | a_0, X) dm$. Note that the parameters ξ , η , and θ are indexed by elements of Q . Thus, a particular choice of Q implies values for each of these parameters as well.

An estimator of $\psi(Q)$ could be constructed by generating estimates \hat{Q} of Q and plugging in:

$$\psi(\hat{Q}) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}(X_i), \quad (\text{plug-in estimator of (1)}) \tag{3}$$

where $\hat{\theta}(x) = \sum_m \hat{\xi}(m, x) \hat{f}_M(m | a_0, x)$ (if M is discrete-valued), $\hat{\xi}(m, x) = \sum_{a=0}^1 \hat{\mu}(m, a, x) \hat{\pi}(a | x)$, and $\hat{\mu}, \hat{\pi}, \hat{f}_M$ are estimates of the outcome regression μ , the propensity score π , and the mediator conditional density f_M , respectively. If M is continuous-valued, then obtaining $\hat{\theta}(x)$ may involve numeric integration (or approximation via Monte Carlo integration under a working model) to compute $\hat{\theta}(x) = \int \hat{\xi}(m, x) \hat{f}_M(m | a_0, x) dm$.

Given a P -integrable function f of the observed data O , let $Pf := \int f(o) p(o) do$ and $P_n f := \frac{1}{n} \sum_{i=1}^n f(O_i)$. A linear expansion of $\psi(\hat{Q})$ yields $\psi(\hat{Q}) = \psi(Q) - P\Phi(\hat{Q}) + R_2(\hat{Q}, Q)$, where Φ is a gradient of ψ satisfying $P\Phi(Q) = 0$, and $R_2(\hat{Q}, Q)$ denotes a second-order remainder term. While multiple gradients may satisfy the expansion, the tangent space of our model is saturated such that there is only a single, unique gradient; known as the efficient influence function (EIF) due to its foundational link to the theory of regular, asymptotically linear estimators [Bickel et al., 1993].

The EIF for $\psi(Q)$ in (1) was provided by Fulcher et al. [2019] and can be written as a sum of four components (see Appendix B.3 for a proof)

$$\begin{aligned} \Phi(Q)(O_i) &= \underbrace{\frac{f_M(M_i | a_0, X_i)}{f_M(M_i | A_i, X_i)} \{Y_i - \mu(M_i, A_i, X_i)\}}_{\Phi_Y(Q)(O_i)} + \underbrace{\frac{\mathbb{I}(A_i = a_0)}{\pi(a_0 | X_i)} \{\xi(M_i, X_i) - \theta(X_i)\}}_{\Phi_M(Q)(O_i)} \\ &\quad + \underbrace{\{\eta(1, X_i) - \eta(0, X_i)\} \{A_i - \pi(1 | X_i)\}}_{\Phi_A(Q)(O_i)} + \underbrace{\theta(X_i) - \psi(Q)}_{\Phi_X(Q)(O_i)}. \end{aligned} \quad (4)$$

For our later use, we note that if M is binary, $\Phi_M(Q)$ can be rewritten (see Appendix B.3),

$$\Phi_M(Q)(O_i) = \frac{\mathbb{I}(A_i = a_0)}{\pi(a_0 | X_i)} \{\xi(1, X_i) - \xi(0, X_i)\} \{M_i - f_M(1 | a_0, X_i)\}. \quad (5)$$

The first-order bias of the plug-in estimator is $-P_n \Phi(\hat{Q})$ (see Appendix B.4). When flexible nuisance estimation strategies are used (e.g., based on machine learning), this term may not have standard root- n asymptotic behavior. This motivates the one-step corrected plug-in estimator, denoted by $\psi_1^+(\hat{Q})$, to be $\psi(\hat{Q}) + P_n \Phi(\hat{Q})$. The one-step estimator takes the form:

$$\psi_1^+(\hat{Q}) = \frac{1}{n} \sum_{i=1}^n \frac{\hat{f}_M(M_i | a_0, X_i)}{\hat{f}_M(M_i | A_i, X_i)} \{Y_i - \hat{\mu}(M_i, A_i, X_i)\} + \frac{\mathbb{I}(A_i = a_0)}{\hat{\pi}(a_0 | X_i)} \{\hat{\xi}(M_i, X_i) - \hat{\theta}(X_i)\} + \hat{\eta}(A_i, X_i), \quad (6)$$

where $\hat{\eta}(a, x) = \int \hat{\mu}(m, a, x) \hat{f}_M(m | a_0, x) dm$.

Fulcher et al. [2019] showed, under parametric working models, this estimator is both asymptotically normal and *doubly robust*, meaning it is consistent for $\psi(Q)$ if either $(\hat{\mu}, \hat{\pi})$ or \hat{f}_M are consistent for their respective target parameters. However, this estimator requires estimating f_M ,

which may be high-dimensional. A second limitation of the one-step approach is its potential to produce estimates outside the parameter space, particularly problematic for binary or bounded outcomes. These drawbacks motivate the development of alternative estimators, such as TMLEs which combine statistical efficiency with guaranteed respect for parameter constraints. Recent work by [Wen et al. \[2024\]](#) addresses some of these concerns. But, notably, their target estimand differs slightly from the standard front-door functional in (1), as they marginalize over the treatment variable early in the derivation, resulting in a decomposition that includes a direct plug-in component and a modified front-door term. While this alternative formulation is well-justified, its statistical structure and interpretation differ from the estimands considered here. Moreover, they do not establish detailed conditions under which flexible learning yields valid inference.

Next, we extend the prior ATE front-door estimation framework by proposing several novel doubly/multiply robust one-step estimators and TMLEs, designed to address the limitations discussed above through flexible nuisance estimation and targeted learning. We further derive novel estimators for the ATT front-door functional—a setting for which no prior estimation methods have been formally proposed.

3 Proposed estimators for the ATE front-door functional

In this section, we present three representations of the EIF for the ATE front-door functional in (1), each tied to a different parameterization of the observed data distribution and motivating distinct estimators. The *first* approach uses the standard factorization and requires direct estimation of all components, including conditional densities (Section 3.1). The *second* and *third* avoid direct estimation of the mediator density by leveraging density ratio reparameterizations or regression-based alternatives (Section 3.2). For each approach, we describe the relevant nuisance components and develop corresponding one-step estimators and TMLEs.

The TMLE construction starts from an initial plug-in estimate $\psi(\hat{Q})$, and updates \hat{Q} to yield \hat{Q}^* by simultaneously (i) reducing empirical risk relative to \hat{Q} and (ii) solving the approximate-equation-solving property where $P_n \Phi(\hat{Q}^*) = o_p(n^{-1/2})$. Concretely, for each nuisance $Q_j \in Q$ we posit a one-dimensional submodel through \hat{Q}_j with an associated loss whose score recovers the corresponding EIF component. Iterative minimization along these submodels yields \hat{Q}^* , and the

final TMLE is $\psi(\hat{Q}^*)$. For details see Appendix B.5 and van der Laan et al. [2011].

Throughout, we assume Y is continuous and defer binary-outcome extensions to Appendix C.2.

3.1 Estimation based on standard factorization

Consider the plug-in estimator in (3), where $Q = (\mu, f_M, \pi, p_X)$ denotes the nuisance functions under the standard factorization of P . The one-step estimator under this parameterization was reviewed in Section 2.2; here, we describe a corresponding TMLE. We begin by obtaining initial estimates $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_X)$. The functions μ and π can be estimated via regression, including machine learning methods, while \hat{p}_X is taken as the empirical distribution of X . The strategy for estimating f_M depends on the nature of the mediator. Here, we focus on direct estimation of the mediator density, which is most practical when M is low-dimensional or discrete. For discrete mediators, standard categorical regression suffices; for continuous, low-dimensional mediators, one may use conditional density estimators ranging from simple parametric models to flexible approaches such as kernel methods or the highly adaptive LASSO [Hayfield and Racine, 2008, Benkeser and Van Der Laan, 2016].

Given an initial estimate \hat{Q} , we outline the targeting step of the TMLE. We begin with binary M and later extend the procedure to accommodate continuous mediators. We assume Q belongs in a functional space \mathcal{Q} , defined as the Cartesian product of each nuisance-functional space \mathcal{M}_{Q_j} .

Binary M . Let $\hat{Q}^{(t)} = (\hat{\mu}^{(t)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$ denote the nuisance estimates at iteration t , with initialization $\hat{Q}^{(0)} = \hat{Q}$. Since the empirical distribution of X satisfies $P_n \Phi_X(\hat{Q}^*) = o_p(n^{-1/2})$, there is no targeting of \hat{p}_X . We therefore focus on updating $\hat{Q}^{(t)} = (\hat{\mu}^{(t)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)})$ to ensure that $P_n \Phi_Y(\hat{Q}^*)$, $P_n \Phi_M(\hat{Q}^*)$, and $P_n \Phi_A(\hat{Q}^*)$ are all $o_p(n^{-1/2})$, where Φ_A and Φ_Y are defined in (4) and Φ_M is given in (5) for binary M . We adopt an iterative procedure with a convergence threshold $C_n = o(n^{-1/2})$, repeating steps (1–4) while $|P_n \Phi(\hat{Q}^{(t)})| > C_n$.

Step 1: Define loss functions and submodels for $\hat{\pi}^{(t)}$, $\hat{f}_M^{(t)}$, and $\hat{\mu}^{(t)}$, satisfying conditions (C1)–(C3).

For a given $\hat{Q}^{(t)} \in \mathcal{Q}$ and $\varepsilon_A, \varepsilon_M, \varepsilon_Y \in \mathbb{R}$, the parametric submodels are defined as:

$$\begin{aligned}\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)})(1 | X) &= \text{expit} \left\{ \logit\{\hat{\pi}^{(t)}(1 | X)\} + \varepsilon_A \{\hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X)\} \right\}, \\ \hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})(1 | A, X) &= \text{expit} \left\{ \logit\{\hat{f}_M^{(t)}(1 | A, X)\} + \varepsilon_M \frac{\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)}{\hat{\pi}^{(t)}(A | X)} \right\},\end{aligned}$$

$$\hat{\mu}(\varepsilon_Y)(M, A, X) = \hat{\mu}^{(t)}(M, A, X) + \varepsilon_Y, \quad (7)$$

where $\hat{\eta}^{(t)}(a^*, X) = \sum_{m=0}^1 \hat{\mu}^{(0)}(m, a^*, X) \hat{f}_m^{(t)}(a_0, X)$ and $\hat{\xi}^{(t)}(m^*, X) = \sum_{a=0}^1 \hat{\mu}^{(0)}(m^*, a, X) \hat{\pi}^{(t)}(a | X)$, for $a^*, m^* \in \{0, 1\}$. Given $\tilde{\pi} \in \mathcal{M}_\pi$, $\tilde{f}_M \in \mathcal{M}_{f_M}$, $\tilde{\mu} \in \mathcal{M}_\mu$, the loss functions are defined as:

$$\begin{aligned} L_A(\tilde{\pi})(O) &= -\log \tilde{\pi}(A | X), \quad L_M(\tilde{f}_M)(O) = -\mathbb{I}(A = a_0) \log \tilde{f}_M(M | A, X), \\ L_Y(\tilde{\mu}; \hat{f}_M^{(t)})(O) &= \{\hat{f}_M^{(t)}(M | a_0, X) / \hat{f}_M^{(t)}(M | A, X)\} \{Y - \tilde{\mu}(M, A, X)\}^2. \end{aligned} \quad (8)$$

See Appendix C.1 for a proof of validity of these submodel–loss function pairs under (C1)–(C3).

We also considered targeting $\hat{\mu}$ using the expit submodel proposed by Gruber and van der Laan [2010], in which Y is first rescaled to the unit interval. This nonlinear submodel has been shown to yield more stable estimates in sparse data settings with low Fisher information [Gruber and van der Laan, 2010]. Details are provided in Appendix C.3.

Because the submodel for $\hat{\mu}^{(t)}$ is linear in ε_Y , the quantities $\hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X)$ and $\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)$ depend only on the initial estimate $\hat{\mu}^{(0)}$. Consequently, the submodels $\hat{\pi}(\varepsilon_A; \hat{\mu}^{(t)}, \hat{f}_m^{(t)})$ and $\hat{f}_M(\varepsilon_M; \hat{\mu}^{(t)}, \hat{\pi}^{(t)})$ depend on $\hat{\mu}^{(t)}$ only through $\hat{\mu}^{(0)}$. We emphasize this by rewriting them as $\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)})$ and $\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})$. Moreover, the loss functions for $\tilde{\pi}$ and \tilde{f}_M are independent of $\hat{\mu}^{(t)}$. Therefore, updates to $\hat{\pi}$ and \hat{f}_M can be performed iteratively without involving updated values of $\hat{\mu}$, which can instead be updated in a single step after finalizing \hat{f}_M (due to its appearance in the loss function for $\tilde{\mu}$).

Step 2: Perform iterative risk minimization to obtain $\hat{\pi}^$ and \hat{f}_M^* .*

Step 2a: Update the estimate of π by solving the empirical risk minimization

$$\hat{\varepsilon}_A = \operatorname{argmin}_{\varepsilon_A \in \mathbb{R}} P_n L_A \left(\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)}) \right). \quad (9)$$

This corresponds to fitting a logistic regression without an intercept term:

$$A \sim \text{offset}(\text{logit } \hat{\pi}^{(t)}(1 | X)) + \hat{H}_A^{(t)}(X), \text{ where } \hat{H}_A^{(t)}(X) := \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X).$$

The auxiliary variable $\hat{H}_A^{(t)}(X)$ is often referred to as the “clever covariate.” The coefficient on this covariate corresponds to $\hat{\varepsilon}_A$, the solution to (9). We update $\hat{\pi}^{(t+1)} = \pi(\hat{\varepsilon}_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)})$ and define $\hat{Q}^{(\text{temp})} = (\hat{\mu}, \hat{\pi}^{(t+1)}, \hat{f}_M^{(t)}, \hat{p}_X)$. Condition (C3) then implies $P_n \Phi_A(\hat{Q}^{(\text{temp})}) = o_p(n^{-1/2})$.

Step 2b: Update the estimate of f_M by solving the empirical risk minimization

$$\hat{\varepsilon}_M = \operatorname{argmin}_{\varepsilon_M \in \mathbb{R}} P_n L_M \left(\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t+1)}) \right). \quad (10)$$

This corresponds to fitting a logistic regression without an intercept term:

$$M \sim \text{offset}(\text{logit } \hat{f}_M^{(t)}(1 | a_0, X)) + \hat{H}_M^{(t)}(X), \text{ where } \hat{H}_M^{(t)}(X) := \frac{\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)}{\hat{\pi}^{(t+1)}(a_0 | X)}.$$

The coefficient on the clever covariate $\hat{H}_M^{(t)}(X)$ yields $\hat{\varepsilon}_M$, the solution to (10). Finally, we update $\hat{f}_M^{(t+1)} = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t+1)})$ and let $\hat{Q}^{(t+1)} = (\hat{\mu}^{(0)}, \hat{\pi}^{(t+1)}, \hat{f}_M^{(t+1)}, \hat{\pi}_X)$. Under condition (C3), this ensures $P_n \Phi_M(\hat{Q}^{(t+1)}) = o_p(n^{-1/2})$. We increment t and repeat *Step 2* until convergence.

Multiple iterations are required because updates to one nuisance parameter affect the auxiliary variable used in updating another. For example, while $P_n \Phi_M(\hat{Q}^{(t+1)}) = o_p(n^{-1/2})$, the term $P_n \Phi_A(\hat{Q}^{(t+1)})$ may no longer satisfy this rate, as updating \hat{f}_M changes the auxiliary variable \hat{H}_A , necessitating a new solution to (9). Likewise, updating $\hat{\pi}$ alters \hat{H}_M , requiring re-optimization of (10). This interdependence of updates and auxiliary variables underlies the need for iteration.

Assume convergence at iteration t^* . Let $\hat{\pi}^* = \hat{\pi}^{(t^*)}$, $\hat{f}_M^* = \hat{f}_M^{(t^*)}$, and $\hat{Q}^{(t^*)} = (\hat{\mu}^{(0)}, \hat{\pi}^*, \hat{f}_M^*)$.

Step 3: Perform one-step risk minimization to obtain $\hat{\mu}^$.*

Update the estimate of μ by solving the empirical risk minimization

$$\hat{\varepsilon}_Y = \operatorname{argmin}_{\varepsilon_Y \in \mathbb{R}} P_n L_Y \left(\hat{\mu}(\varepsilon_Y); \hat{f}_M^* \right). \quad (11)$$

This corresponds to fitting a weighted regression:

$$Y \sim \text{offset}(\hat{\mu}^{(0)}) + 1, \text{ with weight} = \hat{f}_M^*(M | a_0, X) / \hat{f}_M^*(M | A, X).$$

The estimated intercept of this model corresponds to $\hat{\varepsilon}_Y$, as a solution to (11). We update $\hat{\mu}^* = \hat{\mu}(\hat{\varepsilon}_Y; \hat{f}_M^*)$ and define $\hat{Q}^* = (\hat{\mu}^*, \hat{\pi}^*, \hat{f}_M^*)$. Condition (C3) then implies $P_n \Phi_Y(\hat{Q}^*) = 0$.

Step 4: Evaluate the plug-in estimator in (3) using the updated nuisance estimates \hat{Q}^ :*

$$\psi_1(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}^*(X_i), \quad (12)$$

where $\hat{\theta}^*(x) = \sum_{m=0}^1 \hat{\xi}^*(m, x) \hat{f}_M^*(m | a_0, x)$ and $\hat{\xi}^*(m, x) = \sum_{a=0}^1 \hat{\mu}^*(m, a, x) \hat{\pi}^*(a | x)$.

Remark 3.1. The iterative updates of $\hat{\pi}$ and \hat{f}_M can be avoided by using the empirical distribution of (A, X) . This ensures that $P_n [\Phi_A(\hat{Q}^*) + \Phi_X(\hat{Q}^*)] = o_p(n^{-1/2})$, leading to the modified TMLE:

$$\psi_{1,\text{mod}}(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \sum_{m=0}^1 \hat{\mu}^*(m, A_i, X_i) \hat{f}_M^*(m | a_0, X_i). \quad (13)$$

Here, \hat{f}_M^* and $\hat{\mu}^*$ are obtained by solving the respective optimization problems in (10) and (11) sequentially, using a flexible estimate of π to compute the auxiliary variable \hat{H}_M . This approach,

however, introduces a potential inconsistency: it combines two estimates of $p(A|X)$ —one implicit in the empirical distribution and another derived from a regression model for $\pi(A|X)$ used in constructing \hat{H}_M . Despite this incompatibility, the discrepancy is typically negligible.

Continuous M . If M is continuous, the TMLE largely mirrors the binary case, but with additional complexities due to f_M being a conditional probability density function. In this case, we propose to use the following submodel,

$$\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})(M | a_0, X) = \hat{f}_M^{(t)}(M | a_0, X) \left\{ 1 + \varepsilon_M \frac{\hat{\xi}^{(t)}(M, X) - \hat{\theta}^{(t)}(X)}{\hat{\pi}^{(t)}(a_0 | X)} \right\}, \quad (14)$$

where $\hat{\xi}^{(t)}(M, X) = \sum_{a=0}^1 \hat{\mu}^{(0)}(M, a, X) \hat{\pi}^{(t)}(a | X)$ and $\hat{\theta}^{(t)}(X) = \int \hat{\xi}^{(t)}(m, X) \hat{f}_M^{(t)}(m | a_0, X) dm$.

To ensure validity as a submodel of \mathcal{M}_{f_M} , the range of ε_M must be restricted. Appendix C.4 provides details, including an alternative submodel that is more general, but leads to increased computational demand to implement.

The empirical risk minimization problem in (10) requires a grid search or other numerical optimization methods. Upon convergence, condition (C3) ensures that $P_n \Phi_M(\hat{Q}^*) = o_p(n^{-1/2})$. The full TMLE procedure for computing $\psi_1(\hat{Q}^*)$ is summarized in Appendix C.6.

The submodel in (14) also extends to multivariate mediators. However, flexibly estimating f_M in high dimensions presents significant theoretical and computational challenges. To mitigate this, we consider alternative strategies that avoid direct estimation of the conditional mediator density.

3.2 Estimation without density modeling

To bypass mediator density estimation we may reinterpret $\theta(X)$ as a quantity estimable via *sequential regression*. Note that $\theta(X) = \mathbb{E}(\xi(M, X) | A = a_0, X)$. This representation suggests an alternative plug-in estimator of the ATE front-door functional in (1). We first generate estimates $\hat{\mu}$ and $\hat{\pi}$, then define the *pseudo-outcome* variable $\hat{\xi}(M_i, X_i) = \sum_{a=0}^1 \hat{\mu}(M_i, a, X_i) \hat{\pi}(a | X_i)$. To estimate θ , we regress the pseudo-outcome on X using only data points where $A_i = a_0$. This replaces the conditional density estimation with a sequential regression task. We denote this estimate of θ via $\hat{\gamma}$ to distinguish it from the one used previously. Finally, the plug-in estimator can be computed by marginalizing $\hat{\gamma}$ over the empirical distribution of X ,

$$\psi_2(\hat{Q}) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}(X_i). \quad (15)$$

To implement a one-step estimator or TMLE using this plug-in formulation, we must still consider f_M , as it enters $\Phi_Y(Q)$ via the *density ratio* $f_M(M|A=a_0, X)/f_M(M|A, X)$, denoted $f_M^r(M, A, X)$. In multivariate settings, estimating this ratio directly is often more tractable than estimating f_M itself. Several flexible methods exist for direct ratio estimation [Sugiyama et al., 2007, Kanamori et al., 2009, Yamada et al., 2013, Sugiyama et al., 2010]. Alternatively, Bayes' theorem yields a reformulation of f_M^r as:

$$f_M^r(M, A, X) = \frac{\lambda(a_0 | X, M)}{\lambda(A | X, M)} \times \frac{\pi(A | X)}{\pi(a_0 | X)}, \quad (16)$$

where $\lambda(a | x, m) := p(A = a | X = x, M = m)$. This representation enables density ratio estimation through binary regressions for λ and π , offering a practical and flexible alternative to direct ratio estimation. It naturally accommodates multivariate mediators and supports a wide range of tools for binary regression, from logistic regression to machine learning. This reparameterization strategy parallels approaches proposed in prior literature on mediation analysis [Zheng and Van Der Laan, 2012, Díaz et al., 2021].

Similarly, we can adopt a sequential regression approach to estimate η . Since $\eta(A, X) = A\kappa_1(X) + (1 - A)\kappa_0(X)$, where $\kappa_a(X) := \mathbb{E}(\mu(M, a, X) | A = a_0, X)$, we compute $\hat{\mu}(M_i, a, X_i)$ for all i and regress this outcome on X using only observations with $A_i = a_0$, yielding $\hat{\kappa}_a$. Repeating this for $a = \{0, 1\}$ gives $\hat{\eta}(A, X) = A\hat{\kappa}_1(X) + (1 - A)\hat{\kappa}_0(X)$.

Let $\hat{Q} = (\hat{\mu}, \hat{\kappa}_a, \hat{f}_M^r, \hat{\pi}, \hat{\gamma}, \hat{p}_X)$ denote the revised set of nuisance estimates, where \hat{f}_M is replaced by components that avoid conditional density estimation. The one-step estimator is

$$\begin{aligned} \psi_2^+(\hat{Q}) &= \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\gamma}(X_i) + \hat{f}_M^r(M_i, A_i, X_i) \{Y_i - \hat{\mu}(M_i, A_i, X_i)\} \right. \\ &\quad \left. + \frac{\mathbb{I}(A_i = a_0)}{\hat{\pi}(a_0 | X_i)} \{\hat{\xi}(M_i, X_i) - \hat{\gamma}(X_i)\} + \{\hat{\kappa}_1(X_i) - \hat{\kappa}_0(X_i)\} \{A_i - \hat{\pi}(1 | X_i)\} \right\}. \end{aligned} \quad (17)$$

To differentiate the two approaches for estimating f_M^r in $\psi_2^+(\hat{Q})$, we define $\psi_{2a}^+(\hat{Q})$ for direct density ratio estimation, and $\psi_{2b}^+(\hat{Q})$ for the regression-based method via $\hat{\lambda}$ and $\hat{\pi}$.

Given \hat{Q} , we next construct a TMLE based on the sequential regression and density ratio parameterization, following the procedure in Section 3.1 with key modifications outlined below.

Submodels and loss functions. The submodel for $\hat{\mu}$ remains linear with corresponding loss $L_Y(\tilde{\mu}; \hat{f}_M^r) = \hat{f}_M^r(M, A, X) \{Y - \tilde{\mu}(M, A, X)\}^2$. The submodel for $\hat{\pi}$ is defined as in (46), indexed

by $\hat{\kappa}_1(X) - \hat{\kappa}_0(X)$, with standard negative log likelihood loss. In addition, we introduce a linear submodel for $\hat{\gamma}$: $\hat{\gamma}(\varepsilon_\gamma)(X) = \hat{\gamma}(X) + \varepsilon_\gamma$, with loss $L_\gamma(\hat{\gamma}; \hat{\pi}, \hat{\xi})(O) = \frac{\mathbb{I}(A=a_0)}{\hat{\pi}(a_0|X)} \{ \hat{\xi}(M, X) - \hat{\gamma}(X) \}^2$.

See Appendix C.1 for a proof of submodel–loss validity under (C1)–(C3).

Targeting steps. We first update $\hat{\mu}$ via weighted least squares regression with weight $\hat{f}_M^r(M, A, X)$ to obtain $\hat{\mu}^*$. Next, using the updated $\hat{\mu}^*$ to recompute $\hat{\kappa}$, we update $\hat{\pi}$ via logistic regression with no intercept and a single covariate $\hat{\kappa}_1(X) - \hat{\kappa}_0(X)$, yielding $\hat{\pi}^*$. Then, using $\hat{\mu}^*$ and $\hat{\pi}^*$, we compute $\hat{\xi}^*(M, X) = \sum_a \hat{\mu}^*(M, a, X) \hat{\pi}^*(a|X)$ and regress it on X (restricted to $A = a_0$) to estimate $\hat{\gamma}$. An update via weighted regression yields $\hat{\gamma}^*$. See more details in Appendix C.5.

Plug-in estimator. Define $\hat{Q}^* = (\hat{\mu}^*, \hat{\kappa}_a, \hat{f}_M^r, \hat{\pi}^*, \hat{\gamma}^*, \hat{p}_X)$, and evaluate

$$\psi_2(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}^*(X_i). \quad (18)$$

The TMLE that avoids mediator density estimation is detailed in Algorithm 3, Appendix C.6.

As in the one-step case, we define TMLEs $\psi_{2a}(\hat{Q}^*)$ via direct ratio estimation of f_M^r and $\psi_{2b}(\hat{Q}^*)$ via regression using $\hat{\lambda}$ and $\hat{\pi}$.

4 Proposed estimators for the ATT front-door functional

As with the ATE, the ATT front-door functional (2) admits two estimation strategies. First, under the standard factorization of P , one can write $\beta(Q) = \int \sum_{a=0}^1 \frac{\mathbb{I}(a=a_1)}{p(a)} \mu(m, a, x) f_M(m|a_0, x) p(a, x) dm dx$, where $a_1 = 1 - a_0$, and construct density-based estimators (plug-in, one-step and TMLE) by estimating $\mu(m, a, x)$ and the mediator density $f_M(m|a_0, x)$ (with $p(a)$ and $p(a, x)$ from their empirical counterparts). Second, one can bypass estimation of f_M via density-ratio or regression reparameterizations. We focus here on these regression-based approaches and defer the density-based constructions to Appendices D.1 and D.2.

Specifically, we rewrite (2) as $\beta(Q) = \int \sum_{a=0}^1 \frac{\mathbb{I}(a=a_1)}{p(a)} \kappa_a(x) p(a, x) dx$, where $\kappa_a(x) = \mathbb{E}(\mu(m, a, x) | A = a_0, x)$. Let $\hat{Q} = (\hat{\mu}, \hat{\kappa}_a, \hat{p}_A, \hat{p}_{AX})$ denote the collection of nuisance estimates. Estimation procedures for $\hat{\mu}$ and $\hat{\kappa}_a$ are described in Section 3, while \hat{p}_A and \hat{p}_{AX} refer to empirical estimates of $p(A)$ and $p(A, X)$, respectively. This yields the following plug-in estimator:

$$\beta(\hat{Q}) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \hat{\kappa}_{a_1}(X_i). \quad (19)$$

We build on this version of the plug-in to derive a one-step corrected estimator and a TMLE. As a first step, we derive the EIF for $\beta(Q)$, denoted $\Phi_\beta(Q)$ (see Appendix B.3 for a proof):

$$\begin{aligned}\Phi_\beta(Q)(O_i) &= \underbrace{\frac{\mathbb{I}(A_i = a_1)}{p_A(a_1)} f_M^r(M_i, A_i, X_i) \{Y_i - \mu(M_i, A_i, X_i)\}}_{\Phi_{\beta;Y}(Q)(O_i)} \\ &\quad + \underbrace{\frac{\mathbb{I}(A_i = a_0)}{p_A(a_1)} \frac{\pi(a_1 | X_i)}{\pi(a_0 | X_i)} \{\mu(M_i, a_1, X_i) - \kappa_{a_1}(X_i)\}}_{\Phi_{\beta;M}(Q)(O_i)} + \underbrace{\frac{\mathbb{I}(A_i = a_1)}{p_A(a_1)} \{\kappa_{a_1}(X_i) - \beta(Q)\}}_{\Phi_{\beta;AX}(Q)(O_i)}.\end{aligned}\quad (20)$$

Given $\hat{Q} = (\hat{\mu}, \hat{\pi}, \hat{f}_M^r, \hat{\kappa}_{a_1}, \hat{p}_A, \hat{p}_{AX})$, the one-step correction of $\beta(\hat{Q})$, denoted by $\beta^+(\hat{Q})$, is

$$\begin{aligned}\beta^+(\hat{Q}) &= \beta(\hat{Q}) + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \hat{f}_M^r(M_i, A_i, X_i) \{Y_i - \hat{\mu}(M_i, A_i, X_i)\} \right. \\ &\quad \left. + \frac{\mathbb{I}(A_i = a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}(a_1 | X_i)}{\hat{\pi}(a_0 | X_i)} \{\hat{\mu}(M_i, a_1, X_i) - \hat{\kappa}_{a_1}(X_i)\} + \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \{\hat{\kappa}_{a_1}(X_i) - \beta(\hat{Q})\} \right\}.\end{aligned}\quad (21)$$

As in the ATE case, \hat{f}_M^r can be estimated either directly or based on (16) using estimates $\hat{\lambda}$ and $\hat{\pi}$. The corresponding one-step estimators are denoted $\beta_a^+(\hat{Q})$ and $\beta_b^+(\hat{Q})$, respectively.

We next describe a TMLE for the plug-in $\beta(\hat{Q})$ in (19), assuming Y is continuous; modifications for binary outcomes mirror those used for the TMLEs of the ATE and are omitted. It suffices for the updated \hat{Q}^* to satisfy $P_n \Phi_{\beta;Y}(\hat{Q}^*) = o_p(n^{-1/2})$ and $P_n \Phi_{\beta;M}(\hat{Q}^*) = o_p(n^{-1/2})$, as the final term $P_n \Phi_{\beta;AX}(\hat{Q}^*)$ vanishes when p_{AX} is estimated empirically. The TMLE updates $\hat{\mu}$ and $\hat{\kappa}_{a_1}$ using a single-step targeting procedure.

To target $\hat{\mu}$, we define a linear submodel (as in Section 3) and minimize the empirical risk:

$$\hat{\varepsilon}_Y = \operatorname{argmin}_{\varepsilon_Y \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \hat{f}_M^r(M_i, a_1, X_i) \{Y_i - \hat{\mu}(\varepsilon_Y)(M_i, a_1, X_i)\}^2. \quad (22)$$

This corresponds to fitting a weighted regression of the outcome on an intercept-only submodel with offset $\hat{\mu}(M, a_1, X)$ and weights proportional to $\frac{\mathbb{I}(A=a_1)}{\hat{p}_A(a_1)} \hat{f}_M^r(M, a_1, X)$. The updated estimate is $\hat{\mu}^*(m, a, x) = \hat{\mu}(m, a, x) + \hat{\varepsilon}_Y$.

Next, we update $\hat{\kappa}_{a_1}$ via a linear submodel $\hat{\kappa}_{a_1}(\varepsilon_\kappa)(x) = \hat{\kappa}_{a_1}(x) + \varepsilon_\kappa$, minimizing:

$$\hat{\varepsilon}_\kappa = \operatorname{argmin}_{\varepsilon_\kappa \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}(a_1 | X_i)}{\hat{\pi}(a_0 | X_i)} \{\hat{\mu}^*(M_i, a_1, X_i) - \hat{\kappa}_{a_1}(\varepsilon_\kappa)(X_i)\}^2. \quad (23)$$

This corresponds to fitting a weighted regression of $\hat{\mu}^*(M, a_1, X)$ on an intercept with offset $\hat{\kappa}_{a_1}(X)$ and weights $\frac{\mathbb{I}(A=a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}(a_1 | X)}{\hat{\pi}(a_0 | X)}$. The updated function is given by $\hat{\kappa}_{a_1}^*(x) = \hat{\kappa}_{a_1}(x) + \hat{\varepsilon}_\kappa$.

The TMLE is then defined as:

$$\beta(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \hat{\kappa}_{a_1}^*(X_i). \quad (24)$$

As above, \hat{f}_M^r may be estimated either directly or via Bayes' rule, yielding TMLEs denoted by $\beta_a(\hat{Q}^*)$ and $\beta_b(\hat{Q}^*)$, respectively.

5 Inference and asymptotic properties

We now establish the asymptotic properties of our estimators, presenting the expansion using TMLE notation with targeted estimates \hat{Q}^* . The same form and remainder bounds apply to one-step estimators, which we omit for brevity. Given a TMLE $\omega(\hat{Q}^*)$ and EIF $\Phi_\omega(Q)$ for a parameter $\omega(Q)$ —either $\psi(Q)$ or $\beta(Q)$ —its linear expansion takes the form:

$$\omega(\hat{Q}^*) - \omega(Q) = P_n \Phi_\omega(Q) - P_n \Phi_\omega(\hat{Q}^*) + (P_n - P)\{\Phi_\omega(\hat{Q}^*) - \Phi_\omega(Q)\} + R_2(\hat{Q}^*, Q). \quad (25)$$

To establish asymptotic linearity, we require the following conditions:

- (A1) *Donsker estimates*: $\Phi_\omega(\hat{Q}^*) - \Phi_\omega(Q)$ falls in a P -Donsker class with probability tending to 1;
- (A2) *$L^2(P)$ -consistent influence function estimates*: $P\{\Phi_\omega(\hat{Q}^*) - \Phi_\omega(Q)\}^2 = o_p(1)$;
- (A3) *Successful targeting of nuisance parameters*: $P_n \Phi_\omega(\hat{Q}^*) = o_p(n^{-1/2})$.

Conditions (A1)–(A2) imply $(P_n - P)\{\Phi_\omega(\hat{Q}^*) - \Phi_\omega(Q)\} = o_p(n^{-1/2})$, so together with (A3), the expansion in (25) yields $\omega(\hat{Q}^*) - \omega(Q) = P_n \Phi_\omega(Q) + R_2(\hat{Q}^*, Q) + o_p(n^{-1/2})$. It remains to characterize R_2 for each estimator, which we do in separate subsections below, followed by the corresponding asymptotic linearity theorems. Note that finite-dimensional parametric models satisfy the Donsker condition (A1) [van der Vaart and Wellner, 2023]. In Section 5.3, we introduce sample splitting to relax (A1) for flexible nuisance estimators.

Throughout, we let $\|f\| = \sqrt{Pf^2}$ denote the $L^2(P)$ -norm of a P -measurable function f .

5.1 ATE front-door functional estimators

5.1.1 $\psi_1(\hat{Q}^*)$: TMLE with standard factorization

Consider the TMLE $\psi_1(\hat{Q}^*)$ from Section 3.1, with $\hat{Q}^* = (\hat{\mu}^*, \hat{f}_M^*, \hat{\pi}^*, \hat{p}_X)$. Under regularity conditions detailed in Appendix E.1.1, the R_2 remainder for $\psi_1(\hat{Q}^*)$ is bounded by:

$$R_2(\hat{Q}^*, Q) \leq C \left\{ \|\hat{f}_M^* - f_M\| \times \|\hat{\mu}^* - \mu\| + \|\hat{f}_M^* - f_M\| \times \|\hat{\pi}^* - \pi\| \right\}, \quad (26)$$

for some constant $C > 0$. The full expression of $R_2(\hat{Q}^*, Q)$ is provided in Appendix E.1.1. This result paves the way for establishing asymptotic linearity of $\psi_1(\hat{Q}^*)$.

Theorem 5.1 (Asymptotic linearity of $\psi_1(\hat{Q}^*)$). *Suppose the nuisance estimates in \hat{Q}^* have the following $L^2(P)$ convergence rates: $\|\hat{\pi}^* - \pi\| = o_P(n^{-\frac{1}{k}})$, $\|\hat{f}_M^* - f_M\| = o_P(n^{-\frac{1}{b}})$, $\|\hat{\mu}^* - \mu\| = o_P(n^{-\frac{1}{q}})$, and that the convergence exponents satisfy:*

$$(A4.1) \quad \frac{1}{b} + \frac{1}{q} \geq \frac{1}{2} \text{ and } \frac{1}{k} + \frac{1}{b} \geq \frac{1}{2}.$$

Under (A1)–(A3), (A4.1), and regularity conditions (outlined in Appendix E.1.1), $\psi_1(\hat{Q}^*)$ is asymptotically linear: $\psi_1(\hat{Q}^*) - \psi(Q) = P_n \Phi(Q) + o_p(n^{-1/2})$, with influence function $\Phi(Q)$.

Condition (A4.1) ensures $R_2(\hat{Q}^*, Q) = o_p(n^{-1/2})$ via the bound in (26). The cross-product structure allows nuisance estimates to converge at slower than root- n rates, thereby allowing for a potentially wider application of flexible machine learning and statistical models than what is possible under the conditions imposed by Fulcher et al. [2019].

An immediate corollary of Theorem 5.1 is that our TMLE inherits the double robustness properties of the one-step estimator proposed by Fulcher et al. [2019]. While their formulation is framed in terms of parametric working models, we restate the result using $L^2(P)$ -consistency for parsimony and alignment with the TMLEs below.

Corollary 5.2 (Robustness of $\psi_1(\hat{Q}^*)$). $\psi_1(\hat{Q}^*)$ is consistent for $\psi(Q)$ if either (i) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\mu}^* - \mu\| = o_p(1)$, or (ii) $\|\hat{f}_M^* - f_M\| = o_p(1)$, or both (i) and (ii) hold.

5.1.2 $\psi_{2a}(\hat{Q}^*)$: TMLE with direct density ratio and sequential regression

Consider the TMLE $\psi_{2a}(\hat{Q}^*)$ from Section 3.2, where \hat{f}_M^r is obtained via direct density ratio estimation; thus $\hat{Q}^* = (\hat{\mu}^*, \hat{\kappa}_a, \hat{f}_M^r, \hat{\pi}^*, \hat{\gamma}^*, \hat{p}_X)$. Under the regularity conditions detailed in Appendix E.1.2, the R_2 remainder admits the bound:

$$R_2(\hat{Q}^*, Q) \leq C \left\{ \|\hat{f}_M^r - f_M^r\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\pi}^* - \pi\| \times \left\{ \|\hat{\gamma}^* - \gamma\| + \sum_{a=0}^1 \|\hat{\kappa}_a - \kappa_a\| \right\} \right\}, \quad (27)$$

for some finite constant $C > 0$. See the detailed form of $R_2(\hat{Q}^*, Q)$ in Appendix E.1.2. We have the following theorem establishing the asymptotic linearity of $\psi_{2a}(\hat{Q}^*)$.

Theorem 5.3 (Asymptotic linearity of $\psi_{2a}(\hat{Q}^*)$). Suppose the nuisance estimates in \hat{Q}^* satisfy the following $L^2(P)$ convergence rates: $\|\hat{\pi}^* - \pi\| = o_P(n^{-\frac{1}{k}})$, $\|\hat{\mu}^* - \mu\| = o_P(n^{-\frac{1}{q}})$, $\|\hat{\gamma}^* - \gamma\| = o_P(n^{-\frac{1}{j}})$, $\|\hat{\kappa}_a - \kappa_a\| = o_P(n^{-\frac{1}{\ell}})$, $\|\hat{f}_M^r - f_M^r\| = o_P(n^{-\frac{1}{c}})$, and the exponents satisfy:

$$(A4.2) \quad \frac{1}{c} + \frac{1}{q} \geq \frac{1}{2}, \quad \frac{1}{k} + \frac{1}{j} \geq \frac{1}{2}, \text{ and } \frac{1}{\ell} + \frac{1}{k} \geq \frac{1}{2}.$$

Under (A1)-(A3), (A4.2), and regularity conditions (outlined in Appendix E.1.2), $\psi_{2a}(\hat{Q}^*)$ is asymptotically linear: $\psi_{2a}(\hat{Q}^*) - \psi(Q) = P_n \Phi(Q) + o_P(n^{-1/2})$, with influence function $\Phi(Q)$.

$\psi_{2a}(\hat{Q}^*)$ also exhibits multiple robustness.

Corollary 5.4 (Robustness of $\psi_{2a}(\hat{Q}^*)$). $\psi_{2a}(\hat{Q}^*)$ is consistent for $\psi(Q)$ if at least one of the following conditions hold: (i) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\mu}^* - \mu\| = o_p(1)$, (ii) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{f}_M^r - f_M^r\| = o_p(1)$, (iii) $\|\hat{\mu}^* - \mu\| = o_p(1)$, $\|\hat{\gamma}^* - \gamma\| = o_p(1)$, and $\|\hat{\kappa}_a - \kappa_a\| = o_p(1)$, (iv) $\|\hat{\gamma}^* - \gamma\| = o_p(1)$, $\|\hat{\kappa}_a - \kappa_a\| = o_p(1)$, and $\|\hat{f}_M^r - f_M^r\| = o_p(1)$.

Corollary 5.4 highlights that consistency can be achieved either by consistently estimating (μ, π) , or by consistently estimating (γ, κ_a) , and f_M^r . In a partially specified scenario where only one of $\hat{\mu}^*$ or $\hat{\pi}^*$ is consistent, consistency of the estimator still holds if a subset of components in (γ, κ_a) , and f_M^r) is consistently estimated.

5.1.3 $\psi_{2b}(\hat{Q}^*)$: TMLE with fully regression-based methods

Consider the TMLE $\psi_{2b}(\hat{Q}^*)$ from Section 3.2, where f_M^r is estimated via regression-based components π and λ ; thus $\hat{Q}^* = (\hat{\mu}^*, \hat{\kappa}_a, \hat{\lambda}, \hat{\pi}^*, \hat{\gamma}^*, \hat{p}_X)$. Under regularity conditions stated in Appendix E.1.3, the $R_2(\hat{Q}^*, Q)$ term admits the following upper bound

$$C \left\{ \|\hat{\lambda} - \lambda\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\pi}^* - \pi\| \times \{ \|\hat{\mu}^* - \mu\| + \|\hat{\gamma}^* - \gamma\| + \|(\hat{\kappa}_1 - \hat{\kappa}_0) - (\kappa_1 - \kappa_0)\| \} \right\}, \quad (28)$$

for some finite constant $C > 0$. The detailed form of $R_2(\hat{Q}^*, Q)$ is provided in Appendix E.1.3.

Theorem 5.5 (Asymptotic linearity of $\psi_{2b}(\hat{Q}^*)$). Suppose the nuisance estimates in \hat{Q}^* satisfy the following $L^2(P)$ convergence rates: $\|\hat{\pi}^* - \pi\| = o_P(n^{-\frac{1}{k}})$, $\|\hat{\mu}^* - \mu\| = o_P(n^{-\frac{1}{q}})$, $\|\hat{\gamma}^* - \gamma\| = o_P(n^{-\frac{1}{j}})$, $\|\hat{\kappa}_a - \kappa_a\| = o_P(n^{-\frac{1}{\ell}})$, $\|\hat{\lambda} - \lambda\| = o_P(n^{-\frac{1}{d}})$, and the exponents satisfy:

$$(A4.3) \quad \frac{1}{q} + \frac{1}{k} \geq \frac{1}{2}, \quad \frac{1}{d} + \frac{1}{q} \geq \frac{1}{2}, \quad \frac{1}{k} + \frac{1}{j} \geq \frac{1}{2}, \text{ and } \frac{1}{\ell} + \frac{1}{d} \geq \frac{1}{2}.$$

Under (A1)-(A3), (A4.3), and the regularity conditions (outlined in Appendix E.1.3), $\psi_{2b}(\hat{Q}^*)$ is asymptotically linear $\psi_{2b}(\hat{Q}^*) - \psi(Q) = P_n \Phi(Q) + o_p(n^{-1/2})$, with influence function $\Phi(Q)$.

We note that for $\psi_{2b}(\hat{Q}^*)$, consistency of the estimate \hat{f}_M^r depends on both $\hat{\pi}$ and $\hat{\lambda}$, combining robustness conditions (ii) and (iv) from Corollary 5.4. Robustness properties are formalized below.

Corollary 5.6 (Robustness of $\psi_{2b}(\hat{Q}^*)$). $\psi_{2b}(\hat{Q}^*)$ is consistent for $\psi(Q)$ if at least one of the following holds: (i) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\mu}^* - \mu\| = o_p(1)$, (ii) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\lambda} - \lambda\| = o_p(1)$, (iii) $\|\hat{\mu}^* - \mu\| = o_p(1)$, $\|\hat{\gamma}^* - \gamma\| = o_p(1)$, and $\|\hat{\kappa}_a - \kappa_a\| = o_p(1)$.

Unlike ψ_1 and ψ_{2a} , where certain components could ensure consistency on their own, ψ_{2b} requires at least one of $\hat{\mu}^*$ or $\hat{\pi}^*$ to be consistent even when all auxiliary regressions ($\hat{\lambda}, \hat{\gamma}, \hat{\kappa}_a$) are consistently estimated. In this sense, ψ_{2b} exhibits a slightly weaker robustness property. Nevertheless, it remains attractive in practice due to its fully regression-based construction.

5.2 ATT front-door functional estimators

We now establish conditions for the asymptotic linearity of our ATT estimators. Following Section 4, we focus on the fully regression-based estimator $\beta_b(\hat{Q}^*)$, with $\hat{Q}^* = (\hat{\mu}^*, \hat{\kappa}_{a_1}^*, \hat{\lambda}, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$. Under the regularity conditions detailed in Appendix E.2.3, the remainder is bounded by

$$R_2(\hat{Q}^*, Q) \leq C \left\{ \|\hat{\pi} - \pi\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\lambda} - \lambda\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\pi} - \pi\| \times \|\hat{\kappa}_{a_1} - \kappa_{a_1}\| \right\}, \quad (29)$$

for some constant $C > 0$. The detailed form is provided in Appendix E.2.3. Results for $\beta_1(\hat{Q}^*)$ (Appendix D.1) and $\beta_a(\hat{Q}^*)$ (Section 4) are deferred to Appendices E.2.1 and E.2.2, respectively.

Theorem 5.7 (Asymptotic linearity of $\beta_b(\hat{Q}^*)$). Suppose the nuisance estimates in \hat{Q}^* satisfy the following $L^2(P)$ convergence rates: $\|\hat{\pi} - \pi\| = o_p(n^{-\frac{1}{k}})$, $\|\hat{\mu}^* - \mu\| = o_p(n^{-\frac{1}{q}})$, $\|\hat{\kappa}_{a_1} - \kappa_{a_1}\| = o_p(n^{-\frac{1}{d}})$, $\|\hat{\lambda} - \lambda\| = o_p(n^{-\frac{1}{\ell}})$, and the exponents satisfy:

$$(A4.4) \quad \frac{1}{q} + \frac{1}{k} \geq \frac{1}{2}, \quad \frac{1}{d} + \frac{1}{q} \geq \frac{1}{2}, \quad \text{and} \quad \frac{1}{k} + \frac{1}{\ell} \geq \frac{1}{2}.$$

Under (A1)-(A3), (A4.4), and the regularity conditions (outlined in Appendix E.2.3), $\beta_b(\hat{Q}^*)$ is asymptotically linear: $\beta_b(\hat{Q}^*) - \beta_b(Q) = P_n \Phi_\beta(Q) + o_p(n^{-1/2})$, with influence function $\Phi_\beta(Q)$.

Notably, $\beta_b(\hat{Q}^*)$ requires weaker conditions than its ATE counterpart $\psi_{2b}(\hat{Q}^*)$: it avoids the need to estimate γ , which simplifies implementation and strengthens robustness, as shown below.

Corollary 5.8 (Robustness of $\beta_b(\hat{Q}^*)$). $\beta_b(\hat{Q}^*)$ is consistent for $\psi(Q)$ if at least one of the following holds: (i) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\mu}^* - \mu\| = o_p(1)$, (ii) $\|\hat{\pi}^* - \pi\| = o_p(1)$ and $\|\hat{\lambda} - \lambda\| = o_p(1)$, (iii) $\|\hat{\mu}^* - \mu\| = o_p(1)$ and $\|\hat{\kappa}_a - \kappa_a\| = o_p(1)$.

These robustness conditions closely resemble those for $\psi_{2b}(\hat{Q}^*)$ in Corollary 5.6, with one key distinction: consistency of $\hat{\gamma}$ is no longer required. This relaxation simplifies condition (iii) while retaining the benefits of a fully regression-based approach.

5.3 Cross fitting as an alternative to Donsker conditions

Our various estimators of the ATE and ATT can be made robust to violations of the Donsker condition by using sample splitting for nuisance parameter estimation, yielding what is commonly referred to as cross-validated TMLE [Zheng and Van Der Laan, 2010] or double/debiased machine learning [Chernozhukov et al., 2017].

To implement cross-fitting, the data are partitioned into K approximately equal, non-overlapping folds indexed by $S_i \in \{1, \dots, K\}$. For each fold k , nuisance parameters Q are estimated on the data excluding fold k , yielding $\hat{Q}^{(-k)}$. These estimates are then used to evaluate the EIF and generate a cross-fitted one-step estimator or TMLE.

For example, the k -th fold version of the one-step estimator ψ_1^+ is:

$$\begin{aligned} \psi_{1,k}^{+,cf}(\hat{Q}^{(-k)}) &= \frac{1}{n_k} \sum_{i:S_i=k} \frac{\hat{f}_M^{(-k)}(M_i | a_0, X_i)}{\hat{f}_M^{(-k)}(M_i | A_i, X_i)} \{Y_i - \hat{\mu}^{(-k)}(M_i, A_i, X_i)\} \\ &\quad + \frac{\mathbb{I}(A_i = a_0)}{\hat{\pi}^{(-k)}(a_0 | X_i)} \{\hat{\xi}^{(-k)}(M_i, X_i) - \hat{\theta}^{(-k)}(X_i)\} \hat{\eta}^{(-k)}(A_i, X_i), \end{aligned} \quad (30)$$

where $\hat{\xi}^{(-k)}$, $\hat{\theta}^{(-k)}$, and $\hat{\eta}^{(-k)}$ are computed as before using the k -specific nuisance estimates. The final cross-fitted one-step estimator averages over all folds: $\psi_1^{+,cf}(\hat{Q}) = \frac{1}{K} \sum_{k=1}^K \psi_{1,k}^{+,cf}(\hat{Q}^{(-k)})$.

For cross-fitted TMLE, the targeting step is performed using fold-specific submodels that share a common fluctuation parameter. For example, to update $\hat{\pi}$ in $\psi_1(\hat{Q}^*)$, we may define for each k :

$$\hat{\pi}^{(-k)}(\varepsilon_A; \hat{\mu}^{(-k)}, \hat{f}_M^{(-k)}) (1 | X) = \text{expit} \left\{ \text{logit}\{\hat{\pi}^{(-k)}(1 | X)\} + \varepsilon_A \{\hat{\eta}^{(-k)}(1, X) - \hat{\eta}^{(-k)}(0, X)\} \right\},$$

and obtain a shared fluctuation parameter $\hat{\varepsilon}_A$ via pooled empirical risk minimization:

$$\hat{\varepsilon}_A = \arg \min_{\varepsilon_A \in \mathbb{R}} \sum_{k=1}^K P_{n,k} L_A(\hat{\pi}^{(-k)}(\varepsilon_A; \hat{\mu}^{(-k)}, \hat{f}_M^{(-k)})),$$

where $P_{n,k}$ is the empirical distribution of the k -th held-out sample. Analogous submodels can be defined for μ and f_M (Section 3.1) to generate a cross-fitted TMLE.

Cross-fitted estimators retain asymptotic linearity under conditions similar to our earlier theorems, without requiring the Donsker condition (A1). We omit formal statements for brevity.

6 Model evaluation and semiparametric efficiency gains

The assumptions of no unmeasured confounding and no direct effect of A on Y are untestable under the front-door model, which is *nonparametrically saturated* such that it imposes no restrictions on the observed data distribution P . However, [Bhattacharya and Nabi \[2022\]](#) proposed methods for evaluating these assumptions when an *anchor* variable Z is present. An anchor variable is a pre-treatment variable associated with A (and possibly M), but not a direct cause of Y ; i.e., it influences Y only through A and M . In practice, Z can often be viewed as a baseline analogue of the mediator—e.g., pre-vaccine antibody levels when M denotes post-vaccine immune response.

The anchor condition (no direct effect of Z on Y) induces a *generalized independence constraint*—also known as a *Verma* or *dormant* constraint [[Verma and Pearl, 1990](#), [Shpitser and Pearl, 2008](#)]—in P over $O = (X, Z, A, M, Y)$. Such constraints arise as independence relations in truncated or post-intervention distributions. In the anchor-included front-door model, the Verma takes the form $Z \perp Y^m | X$, equivalent to $Z \perp Y$ in the truncated distribution $P(O)/P(M | A, Z, X)$; see Appendix F.1 for details. This constraint underlies the parametric tests of [Bhattacharya and Nabi \[2022\]](#) for assessing the joint validity of conditional ignorability and the absence of a direct effect of A on Y (see their proof of Theorem 1 and Appendix B).

We advance anchor variable testing on three fronts. First, we generalize the prior parametric tests to allow flexible nuisance estimations, e.g., based on modern machine learning, yielding a flexible weighted risk minimization framework (Section 6.1). Second, we introduce a novel *doubly robust* test based on a *conditional counterfactual means*, which remains valid under partial model misspecification and is particularly well-suited to settings where the anchor and mediator are discrete or can be discretized (Section 6.2). Third, we show that when the Verma constraint holds, it can be leveraged to construct more efficient estimators for causal effects (Section 6.3).

6.1 Testing via weighted risk minimizations

The Verma constraint $Z \perp Y^m \mid X$ is equivalently expressed as $Z \perp Y^a \mid X, M^a$ (see Theorem 1 in [Bhattacharya and Nabi, 2022] and Appendix F.1), which implies that, under the null hypothesis that the front-door assumptions hold, the conditional distribution $P(Y^a \mid M^a, Z, X)$ is invariant to Z . Here, we test a specific implication of this constraint: that the conditional mean $\mathbb{E}(Y^a \mid M^a, Z, X)$ should be invariant in Z . While this implication is weaker than full distributional invariance, it suffices for evaluating identification of the causal effects. Under the null, the following MSE risk minimizers coincide (and for binary Y , so do the corresponding distributions):

$$\begin{aligned}\mu_{\text{primal}}^a(m, z, x) &:= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \int (y - \tilde{\mu}(m, z, x))^2 dP(Y^a = y, M^a = m, z, x), \\ \mu_{\text{primal}}^a(m, x) &:= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \int (y - \tilde{\mu}(m, x))^2 dP(Y^a = y, M^a = m, x).\end{aligned}\tag{31}$$

The minimizers in (31) can be re-expressed as weighted risk minimizers under P (see Appendix F.1 for identification details):

$$\begin{aligned}\mu_{\text{primal}}^a(m, z, x) &= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \mathbb{E}(\text{q}_{\text{primal}}(A \mid Y, M, Z, X)(Y - \tilde{\mu}(M, Z, X))^2), \\ \mu_{\text{primal}}^a(m, x) &= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \mathbb{E}(\text{q}_{\text{primal}}(A \mid Y, M, Z, X)(Y - \tilde{\mu}(M, X))^2),\end{aligned}\tag{32}$$

where q_{primal} is the *primal weight* [Bhattacharya et al., 2022], defined as

$$\text{q}_{\text{primal}}(A \mid Y, M, Z, X) = \frac{\sum_{a'} \pi(a' \mid Z, X) f_Y(Y \mid M, a', Z, X)}{\pi(A \mid Z, X) f_Y(Y \mid M, A, Z, X)}.$$

To implement the test via (32), we first estimate q_{primal} using models for the propensity score $\pi(A = a \mid Z, X) := p(A = a \mid Z, X)$ and the conditional outcome density $f_Y(Y \mid M, A, Z, X) := p(Y \mid M, A, Z, X)$. Notably, the outcome density ratio can be estimated via Bayes' rule from $p(A \mid Y, M, Z, X)$ and $p(A \mid M, Z, X)$. Given the estimate $\hat{\text{q}}_{\text{primal}}$, we fit two primal-weighted regressions of Y on (M, Z, X) and (M, X) to estimate the minimizers in (32). We define the *primal test statistic* as the difference in empirical MSE risks:

$$T_{n,\text{primal}} = \frac{1}{n} \sum_{i=1}^n \{(Y_i - \hat{\mu}_{\text{primal}}^a(M_i, Z_i, X_i))^2 - (Y_i - \hat{\mu}_{\text{primal}}^a(M_i, X_i))^2\}.\tag{33}$$

To approximate the null distribution of $T_{n,\text{primal}}$, we adopt a permutation approach [Paschali et al., 2022]. Specifically, we permute the values of Z across observations, refit the two weighted regressions, and recompute the primal test statistic. Repeating this procedure multiple times

yields a reference distribution under the null. The one-sided p-value is computed as the proportion of permuted test statistics greater than or equal to the observed value.

This permutation-based approach remains valid even when regression models are fit using flexible machine learning methods due to the nonparametric nature of the test [Paschali et al., 2022]. Unlike bootstrap procedures—which may break down in non-Donsker settings or yield unstable results with complex learners—the permutation test relies only on the assumption that, under the null, the primal-weighted distribution of Y is invariant to permutations of Z given (M, X) . This form of conditional exchangeability ensures the validity of the test without requiring asymptotic approximations or regularity conditions on the estimators.

The validity of the primal test relies on correct specification of both the treatment and outcome models: π, f_Y . Bhattacharya and Nabi [2022] proposed a complementary parametric test based on the following *dual weight*, which re-weights P using $f_M(M | A, Z, X) := p(M | A, Z, X)$:

$$q_{\text{dual}}(M | A, Z, X) = f_M(M | a, Z, X) / f_M(M | A, Z, X).$$

The counterfactual risk minimizations in (31) can be implemented via weighted least squares using q_{dual} (see Appendix F.1 for a proof.) Consequently, replacing q_{primal} with q_{dual} in (32) yields a nonparametric dual test. To implement it, we first estimate q_{dual} (e.g., via density-ratio estimation or Bayes-rule decomposition). With the resulting estimate \hat{q}_{dual} , we fit two weighted regressions of Y on (M, Z, X) and (M, X) , yielding estimates $\hat{\mu}_{\text{dual}}^a(M, Z, X)$ and $\hat{\mu}_{\text{dual}}^a(M, X)$, respectively. The *dual test statistic* is defined analogously to the primal case:

$$T_{n,\text{dual}} = \frac{1}{n} \sum_{i=1}^n \{(Y_i - \hat{\mu}_{\text{dual}}^a(M_i, Z_i, X_i))^2 - (Y_i - \hat{\mu}_{\text{dual}}^a(M_i, X_i))^2\}. \quad (34)$$

As in the primal test, we approximate the null distribution of $T_{n,\text{dual}}$ using a permutation procedure. We repeatedly permute the values of Z , refit the weighted regressions, and recalculate the test statistic. The one-sided p-value is defined as the proportion of permuted statistics less than or equal to the observed $T_{n,\text{dual}}$. This approach supports flexible or nonparametric regressions while maintaining valid inference under the null.

While the primal and dual tests offer complementary strengths—the former relying on treatment and outcome models, the latter on the mediator model—each requires correct specification of at least one set of nuisance components. In practice, model misspecification can undermine the

validity of either test, and conflicting results may be difficult to interpret. This motivates our next *doubly robust* test based on the invariance of a conditional counterfactual mean (CCM).

6.2 A doubly robust test

We assume M and Z are discrete (or discretized), deferring continuous-valued cases to future work. Under the Verma $Z \perp Y^m | X$, we have $\mu^m(z, x) := \mathbb{E}(Y^m | Z = z, X = x)$ constant in z , for every (m, x) . When X is discrete and low-dimensional, one can test pointwise invariance by checking $\mu^m(1, x) = \mu^m(0, x)$ within each stratum of X via a Wald-type test (see Appendix F.2). However, if X is continuous or high-dimensional, this approach is not feasible due to the curse-of-dimensionality. In this instance, we suggest that a test could be based on the marginalized quantity $\mu^m(z) := \int \mu^m(z, x) p(x) dx$, and test a weaker null: $\Delta(m) := \mu^m(1) - \mu^m(0) = 0$. This test has the advantage of being based on a pathwise differentiable parameter $\Delta(m)$, allowing the utilization of doubly robust methods, as described below. However, depending on the structure of $\mu^m(z, x)$, it may have limited power against some alternatives. Nevertheless, characterizing a robust test based on the marginal parameter $\Delta(m)$ may prove useful in many settings.

Let Δ_n denote a vector of estimated contrasts $\Delta_n(m)$ for each m . Let Σ_n denote an estimate of the asymptotic variance-covariance matrix of $n^{1/2}\Delta_n$, which can generally be obtained as the empirical covariance matrix of estimated influence functions. A Wald-style test statistic is defined as $T_{n,CCM} := \Delta_n^\top \Sigma_n^{-1} \Delta_n / n$. Under the null and in large samples, $T_{n,CCM}$ is approximately Chi-squared distributed with d degrees of freedom, where d is the dimension of Δ . Comparison of the test statistic to relevant quantiles of this distribution allows for appropriate hypothesis tests with correct asymptotic size.

To implement this test, we require robust estimates of both Δ and the covariance matrix Σ . Estimators of Δ are motivated by the identification result that (see Appendix F.2 for proof)

$$\mu^m(z) = \int \sum_a \mu(m, a, z, x) \pi(a | z, x) p(x) dx . \quad (35)$$

Plug-in estimators based on (35) may suffer from the *g-null paradox* [Robins and Wasserman, 1997], whereby parametric estimation of both μ and π can lead to invalid tests that reject the null even when it holds. This motivates the usage of influence-function-based estimators that remain valid under flexible nonparametric estimation of nuisance components—even when convergence rates

fall below root- n . For example, a one-step estimator of $\mu^m(z)$ can be computed as follows. We define $f_Z(Z|X) := p(Z|X)$ and propose the estimator (see detailed derivation in Appendix F.2):

$$\begin{aligned}\hat{\mu}^{+,m}(z) = & \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(Z_i = z, M_i = m)}{\hat{f}_M(m | A_i, z, X_i) \hat{f}_Z(z | X_i)} (Y_i - \hat{\mu}(m, A_i, z, X_i)) + \sum_a \hat{\mu}(m, a, z, X_i) \hat{\pi}(a | z, X_i) \\ & + \frac{\mathbb{I}(Z_i = z)}{\hat{f}_Z(z | X_i)} (\hat{\mu}(m, A_i, z, X_i) - \sum_a \hat{\mu}(m, a, z, X_i) \hat{\pi}(a | z, X_i)).\end{aligned}\quad (36)$$

The above estimator, and the TMLE counterpart [Gruber and van der Laan, 2010], exhibit doubly-robust consistency for $\mu^m(z)$ if either $(\hat{\pi}, \hat{\mu})$ or (\hat{f}_M, \hat{f}_Z) are consistent.

While doubly-robust estimation of $\mu^m(z)$ (and thereby Δ) is straightforward, ensuring a doubly-robust estimate of the variance-covariance matrix Σ is more challenging. The challenge arises from the fact that under inconsistent estimation of nuisance parameters, the one-step (TMLE) estimate of $\mu^m(z)$, while consistent, will not generally be asymptotically linear, unless it is based on working parametric models. However, as noted above their use in this case is susceptible to the g-null paradox and therefore is not recommended. If flexible regressions with slower-than-parametric convergence rates are adopted, then additional effort is required to ensure doubly robust asymptotic linearity [Van der Laan, 2014] and generally this has only been demonstrated to be feasible using TMLE [Benkeser et al., 2017].

Thus, we propose to adopt these TMLE-based methods to develop a doubly robust hypothesis test. This involves a careful analysis of the second-order remainder term (see Appendix F.2 for details). We refer to this test as DR-CCM.

The three tests offer flexible validation of the front-door model. DR-CCM is doubly robust but limited to discrete mediators/anchors; the primal test handles continuous or multivariate settings under correct treatment and outcome models; and the dual test only requires a correct mediator model. The test should be based on which nuisance component can be most reliably estimated.

6.3 Efficiency gains under the Verma constraint

When the Verma constraint holds (i.e., under the null), it imposes structural restrictions on the observed data distribution, shrinking the statistical model and enabling the construction of more efficient estimation of causal effects. We illustrate this in the context of estimating $\mathbb{E}(Y^{a_0})$.

Under the front-door model with an anchor variable Z , we define a family of identification

functionals for $\mathbb{E}(Y^{a_0})$, each indexed by a fixed level z^* in the state space \mathcal{Z} of Z :

$$\psi_{z^*}(\mathbf{Q}) = \iiint \sum_{a=0}^1 \mu(m, a, z^*, x) \pi(a | z^*, x) f_M(m | A = a_0, z, x) p(z, x) dm dz dx. \quad (37)$$

See Appendix F.3 for an identification proof. Although $\psi_{z^*}(\mathbf{Q})$ equals $\mathbb{E}(Y^{a_0})$ for all $z^* \in \mathcal{Z}$, the efficiency of plug-in or influence-function-based estimators may vary with the choice of z^* . Below, we focus on one-step estimators that avoid density estimation and show how to exploit this structure to improve efficiency, beginning with the case where Z is discrete.

Estimation under discrete Z . A one-step estimator for (37) can be constructed using this set of nuisance functions: $p_{ZX}(z, x) := p(Z = z, X = x)$, $f_Z(z | x)$, $\pi(a | z, x)$, $\mu(m, a, z, x)$, $\xi_{z^*}(m, x) := \sum_a \mu(m, a, z^*, x) \pi(a | z^*, x)$, $\gamma_{z^*}(z, x) := \mathbb{E}(\xi_{z^*}(M, X) | a_0, z, x)$, $\kappa_{a,z^*}(z, x) := \mathbb{E}(\mu(M, a, z^*, X) | a_0, z, x)$, and $f_{M,z^*}^r(m, a, z, x) := f_M(m | a_0, z, x) / f_M(m | a, z^*, x)$. Let $\mathbf{Q} = \{\mu, \pi, \xi_{z^*}, \gamma_{z^*}, \kappa_{a,z^*}, f_{M,z^*}^r, f_Z, p_{ZX}\}$. Given the nuisance estimates, $\hat{\mathbf{Q}}$, the one-step estimator is given as $\psi_{z^*}^+(\hat{\mathbf{Q}}) = \frac{1}{n} \sum_{i=1}^n \Phi_{z^*}(\hat{\mathbf{Q}})(O_i) + \hat{\gamma}_{z^*}(Z_i, X_i)$, where $\Phi_{z^*}(\mathbf{Q})$ denotes the np-EIF of (37) and is given by (see a proof in Appendix F.3):

$$\begin{aligned} \Phi_{z^*}(\mathbf{Q})(O_i) &= \frac{\mathbb{I}(Z_i = z^*)}{f_Z(z^* | X_i)} \sum_z f_{M,z^*}^r(M_i, A_i, z, X_i) f_Z(z | X_i) (Y_i - \mu(M_i, A_i, z^*, X_i)) \\ &\quad + \frac{\mathbb{I}(Z = z^*)}{f_Z(z^* | X_i)} (A_i - \pi(a | z^*, X_i)) \sum_z (\kappa_{1,z^*}(z, X_i) - \kappa_{0,z^*}(z, X_i)) f_Z(z | X_i) \\ &\quad + \frac{\mathbb{I}(A_i = a_0)}{\pi(a_0 | Z_i, X_i)} (\xi_{z^*}(M_i, X_i) - \gamma_{z^*}(Z_i, X_i)) + \gamma_{z^*}(Z_i, X_i) - \psi_{z^*}(\mathbf{Q}). \end{aligned} \quad (38)$$

Although the estimand in (37) is invariant to the choice of z^* , the efficiency of the estimator $\psi_{z^*}^+(\hat{\mathbf{Q}})$ generally is not. To explore this, we define a *class of influence functions* formed by convex combinations of the EIFs corresponding to different anchor levels. Under binary Z , this class is

$$\Lambda_\alpha := \{\alpha \Phi_{z^*=1}(\mathbf{Q}) + (1 - \alpha) \Phi_{z^*=0}(\mathbf{Q}), \quad \text{for } \alpha \in [0, 1]\}. \quad (39)$$

For any fixed $\alpha \in \mathbb{R}$ and $\hat{\mathbf{Q}}$, we define the aggregated estimator as $\psi_\alpha^+(\hat{\mathbf{Q}}) := \alpha \psi_{z^*=1}^+(\hat{\mathbf{Q}}) + (1 - \alpha) \psi_{z^*=0}^+(\hat{\mathbf{Q}})$. When $\alpha = 0$ or 1 , this reduces to $\psi_{z^*=0}^+(\hat{\mathbf{Q}})$ or $\psi_{z^*=1}^+(\hat{\mathbf{Q}})$, respectively. To improve efficiency, we derive an *optimal weight* α^{opt} that minimizes the asymptotic variance of the aggregated estimator, given by the variance of the combined influence functions:

$$\alpha^{\text{opt}} := \operatorname{argmin}_{\alpha \in [0, 1]} \mathbb{E}(\{\alpha \Phi_{z^*=1}(\mathbf{Q}) + (1 - \alpha) \Phi_{z^*=0}(\mathbf{Q})\}^2). \quad (40)$$

The minimizer has a closed form: $\alpha^{\text{opt}} = \mathbb{E}(\Phi_{z^*=0}(\mathbf{Q})(\Phi_{z^*=0}(\mathbf{Q}) - \Phi_{z^*=1}(\mathbf{Q}))) / \mathbb{E}((\Phi_{z^*=1}(\mathbf{Q}) -$

$\Phi_{z^*=0}(Q))^2$) (see Appendix F.3 for a proof), which can be estimated using the empirical variances of the influence functions at each level of z^* . The resulting *optimally weighted estimator* is

$$\psi_{\alpha^{\text{opt}}}^+(\hat{Q}) = \hat{\alpha}^{\text{opt}} \psi_{z^*=1}^+(\hat{Q}) + (1 - \hat{\alpha}^{\text{opt}}) \psi_{z^*=0}^+(\hat{Q}).$$

Extension to continuous Z . When Z is continuous, the functional $\psi_{z^*}(Q)$ in (37) is not pathwise differentiable, so a von Mises expansion does not apply. One practical solution is to discretize Z using meaningful cutoffs and apply the discrete methods. Alternatively, one can define an integrated functional by averaging $\psi_{z^*}(Q)$ over a reference distribution $\tilde{p}(Z)$ with the same support as the true marginal of Z (see Appendix F.3):

$$\psi_{\tilde{p}}(Q) = \iiint \left\{ \int \sum_a \mu(m, a, z, x) \pi(a | z, x) \tilde{p}(z) dz \right\} f_M(m | a_0, z, x) p(z, x) dm dz dx. \quad (41)$$

As in the discrete case, the estimand remains invariant to the choice of \tilde{p} , though the efficiency of the resulting estimator may depend on it. Details on constructing one-step estimators based on $\psi_{\tilde{p}}$ and leveraging the Verma constraint in this setting are provided in Appendix F.3.

7 Simulation studies

We conducted six sets of simulation studies, each targeting a distinct methodological question addressed in this paper. (1) *Theoretical properties*: Assessed the asymptotic behavior of the ATE and ATT estimators under various settings, including both uni- and multivariate mediators. This scenario also compared TMLEs using linear versus nonlinear submodels. (2) *Weak overlap*: Examined the potential finite-sample advantages of TMLEs over one-step estimators for both ATE and ATT under weak treatment overlap; (3) *Model misspecification*: Evaluated the robustness of the ATE and ATT estimators when nuisance models were correctly specified versus misspecified; (4) *Cross-fitting*: Investigated whether cross-fitting improves performance for TMLE and one-step estimators of ATE and ATT in settings prone to overfitting; (5) *Model evaluation*: Analyzed type I error and power of our three proposed tests for validity of the front-door model assumptions under various null and alternative scenarios; and (6) *Efficiency gain*: Demonstrated that incorporating the Verma constraint within a semiparametric model improves the efficiency of ATE estimation.

ATE was estimated as contrasts of the estimated $\psi_{a_0}(P)$ for $a_0 \in \{0, 1\}$, following Section 3. ATT was estimated by estimating $\beta_{a_0}(P)$ for $a_0 = 0$, following Section 4, and subtracting it from

the empirical mean of Y among individuals with $A = 1$. With slight abuse of notation, we use the same symbols to represent the corresponding contrasts in the ATE and ATT estimators.

The implementation is available in the GitHub repository: [annaguo-bios/fd-methods](#). We have also developed the [fdcausal](#) R package for causal inference under the front-door model.

Simulation 1: Theoretical properties. We assessed asymptotic bias and variance of our ATE and ATT estimators across mediators (binary, univariate to four-dimensional continuous) using parametric and kernel nuisance fits, confirming \sqrt{n} -bias decay and variance convergence to $P[\Phi(Q)^2]$ (Appendix G.1, ATE: Figs (5)–(8); ATT: Figs (9)–(12)). We also compared linear versus expit TMLE submodels on bias, standard deviation (SD), mean squared error (MSE), and 95% confidence interval (CI) coverage and width for select mediators, finding both valid under correct model specification (Appendix G.1, ATE: Table 5; ATT: Table 6).

Simulation 2: Weak overlap. We evaluated TMLE and one-step estimators for ATE and ATT under weak overlap, induced by assigning $A | X = x \sim \text{Bernoulli}(0.001 + 0.998x)$ for $X \sim \text{Uniform}(0, 1)$, yielding near-deterministic probabilities. See Appendix G.2 for details.

We considered three mediator settings: univariate binary, univariate continuous, and bivariate continuous. For each, we implemented practical ATE estimators. In the binary case, we used $\psi_1^+(\hat{Q})$ and $\psi_1(\hat{Q}^*)$, leveraging the ease of modeling binary mediator densities. For continuous mediators, we included $\psi_1^+(\hat{Q})$, $\psi_1(\hat{Q}^*)$, $\psi_{2a}^+(\hat{Q})$, $\psi_{2a}(\hat{Q}^*)$, $\psi_{2b}^+(\hat{Q})$, and $\psi_{2b}(\hat{Q}^*)$. Mediator-related nuisance functions were estimated using kernel density estimation, density-ratio methods, and Bayes-based regression. ATT estimators were constructed analogously.

Based on 1000 replicates at sample sizes of 500, 1000, and 2000, we assessed bias, SD, MSE, CI coverage, and width. ATE results, provided in Table 1, show comparable bias across estimators, but TMLEs had lower SD and narrower CIs, yielding reduced MSE across all mediator types and sample sizes. ATT results appear in Appendix G.2, Table 7.

Simulation 3: Model misspecification. We evaluated the sensitivity of ATE and ATT estimators to model misspecification by comparing parametric models (main terms only) with flexible nuisance estimation via Super Learner—an ensemble of GLMs, GAMs, random forests, SVMs, BART, and XGBoost [Van der Laan et al., 2007]. To address potential Donsker violations from complex learners, we also included cross-fitted versions of all estimators.

Table 1: Comparison of ATE TMLE and one-step estimators under weak overlap across mediator types.

	Univariate Binary				Univariate Continuous				Bivariate Continuous			
	$\psi_1(\hat{Q}^*)$	$\psi_1^+(\hat{Q})$	$\psi_1(\hat{Q}^*)$	$\psi_1^+(\hat{Q})$	$\psi_{2a}(\hat{Q}^*)$	$\psi_{2a}^+(\hat{Q})$	$\psi_{2b}(\hat{Q}^*)$	$\psi_{2b}^+(\hat{Q})$	$\psi_{2a}(\hat{Q}^*)$	$\psi_{2a}^+(\hat{Q})$	$\psi_{2b}(\hat{Q}^*)$	$\psi_{2b}^+(\hat{Q})$
	Bias	SD	MSE	Coverage	CI width	SD	MSE	Coverage	CI width	SD	MSE	Coverage
n=500	-0.004	-0.01	-0.022	-0.004	-0.002	0	-0.002	-0.012	-0.012	0.153	-0.031	-0.065
	0.078	0.418	0.135	0.799	0.432	2.524	0.405	1.191	0.61	5.096	0.495	1.447
	0.006	0.174	0.019	0.638	0.187	6.363	0.164	1.418	0.372	25.965	0.245	2.097
	91.2%	95.4%	96.6%	95.2%	98.4%	97.1%	98.3%	97.3%	99.4%	98.2%	98.5%	97.7%
	0.317	0.854	1.533	1.531	4.764	5.705	2.72	3.447	10.115	12.1	2.854	3.834
n=1000	0	-0.002	-0.012	-0.018	-0.004	0.041	-0.003	0.02	-0.015	-0.078	-0.003	-0.001
	0.056	0.207	0.101	0.47	0.342	1.394	0.338	0.787	0.389	1.841	0.333	0.716
	0.003	0.043	0.01	0.221	0.117	1.942	0.114	0.619	0.152	3.391	0.111	0.513
	92.1%	95.4%	96%	94.3%	98.5%	96.3%	98%	97.1%	99.4%	97.1%	99%	96.4%
	0.24	0.492	0.931	0.93	3.071	3.46	1.861	2.178	4.809	5.365	1.852	2.136
n=2000	0	-0.002	-0.005	0.01	0.009	0.01	0.009	0.014	0.003	-0.006	0.008	0.022
	0.039	0.114	0.068	0.239	0.238	0.699	0.243	0.481	0.319	0.98	0.276	0.489
	0.001	0.013	0.005	0.057	0.057	0.488	0.059	0.231	0.102	0.959	0.076	0.24
	94.1%	96.2%	97.4%	96%	99.2%	96.9%	98.7%	96%	99.2%	96.9%	98.6%	97.4%
	0.175	0.318	0.602	0.602	1.96	2.092	1.321	1.454	2.989	3.209	1.351	1.504

Simulations used binary and continuous mediators, 1000 replicates, and sample sizes of 500, 1000, and 2000 (details in Appendix G.3). For binary mediators, we used $\psi_1^+(\hat{Q})$ and $\psi_1(\hat{Q}^*)$; for continuous mediators, $\psi_{2a}(\hat{Q}^*)$, $\psi_{2b}(\hat{Q}^*)$, and their one-step analogues. ATE results, provided in Table 2, show that misspecified models led to bias and poor coverage, while Super Learner-based estimators reduced bias and improved coverage with increasing sample size. Some undercoverage persisted for ψ_1 , and cross-fitting yielded limited additional gains. These results highlight the importance of flexible nuisance estimation. ATT findings (Appendix G.3, Table 8) were similar.

Simulation 4: Cross-fitting. We examined the role of cross-fitting by focusing on random forests, which are known to perform poorly without sample splitting in high-dimensional settings [Chernozhukov et al., 2017, Biau, 2012]. Details are provided in Appendix G.4 (see Tables 9-12).

Simulation 5: Model evaluation. We evaluated the performance of the proposed tests from Section 6 using simulations designed to assess type I error and power. Each scenario involved 200 replicates per sample size, with the rejection rate interpreted as type I error when the data-generating process satisfied front-door assumptions, and as power when it did not. We used four data-generating models: in DAG1, Z has direct effects on both A and M ; in DAG2, Z affects A and shares unmeasured confounding with M —both satisfying the front-door conditions.

Table 2: Performance of ATE estimators under model misspecifications across mediator types.

	TMLEs												One-step estimators												
	Univariate Binary						Univariate Continuous						Univariate Binary						Univariate Continuous						
	$\psi_1(\hat{Q}^*)$			$\psi_{2a}(\hat{Q}^*)$			$\psi_{2b}(\hat{Q}^*)$			$\psi_1^+(\hat{Q})$			$\psi_{2a}^+(\hat{Q})$			$\psi_{2b}^+(\hat{Q})$									
	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	
n=500	Bias	-0.016	-0.001	-0.01	-0.081	-0.02	-0.037	-0.081	-0.016	-0.038	-0.017	-0.008	-0.005	-0.081	-0.021	-0.039	-0.081	-0.016	-0.037						
	SD	0.043	0.05	0.071	0.099	0.123	0.128	0.099	0.116	0.123	0.043	0.048	0.183	0.099	0.128	0.133	0.099	0.115	0.126						
	MSE	0.002	0.003	0.005	0.016	0.016	0.018	0.016	0.014	0.016	0.002	0.002	0.033	0.016	0.017	0.019	0.016	0.014	0.017						
	Coverage	84.2%	83.2%	82.8%	85.5%	97%	96.8%	85.5%	91.5%	91.8%	83.1%	80%	81.5%	85.5%	96.8%	96.5%	85.5%	91.4%	91.4%						
	CI width	0.161	0.154	0.172	0.398	0.567	0.596	0.399	0.398	0.444	0.158	0.143	0.176	0.399	0.56	0.589	0.399	0.397	0.444						
	Bias	-0.018	-0.003	-0.008	-0.081	-0.012	-0.027	-0.081	-0.009	-0.023	-0.018	-0.006	-0.008	-0.081	-0.013	-0.029	-0.081	-0.009	-0.023						
n=1000	SD	0.03	0.035	0.035	0.074	0.088	0.089	0.074	0.088	0.089	0.03	0.034	0.035	0.074	0.092	0.092	0.074	0.087	0.089						
	MSE	0.001	0.001	0.001	0.012	0.008	0.009	0.012	0.008	0.008	0.001	0.001	0.001	0.012	0.009	0.009	0.012	0.008	0.008						
	Coverage	81.5%	87.3%	85.3%	74.6%	98.2%	97.2%	74.6%	90.1%	89.9%	80.8%	83.6%	84.2%	74.6%	96.8%	96.6%	74.6%	90.3%	89.8%						
	CI width	0.111	0.113	0.117	0.282	0.403	0.416	0.282	0.293	0.311	0.109	0.106	0.11	0.282	0.4	0.412	0.282	0.292	0.31						
	Bias	-0.018	-0.002	-0.005	-0.084	-0.008	-0.019	-0.084	-0.005	-0.016	-0.018	-0.004	-0.005	-0.084	-0.008	-0.018	-0.084	-0.005	-0.016						
	SD	0.02	0.023	0.024	0.05	0.06	0.059	0.05	0.06	0.059	0.02	0.023	0.023	0.05	0.062	0.061	0.05	0.06	0.059						
n=2000	MSE	0.001	0.001	0.001	0.01	0.004	0.004	0.01	0.004	0.004	0.001	0.001	0.001	0.01	0.004	0.004	0.01	0.004	0.004						
	Coverage	76.9%	89.7%	88.4%	60.5%	97.9%	98%	60.4%	92.2%	92.5%	75.4%	87.2%	87.4%	60.5%	97.3%	97.6%	60.4%	92.1%	92.3%						
	CI width	0.077	0.083	0.084	0.198	0.288	0.293	0.198	0.214	0.222	0.076	0.079	0.081	0.198	0.286	0.291	0.198	0.213	0.221						

Table 3: Comparative analysis of DR-CCM, dual, and primal tests under model misspecifications.

n	DR-CCM test				Dual test				Primal test			
	Type I error		Power		Type I error		Power		Type I error		Power	
	DAG1	DAG2	DAG3	DAG4	DAG1	DAG2	DAG3	DAG4	DAG1	DAG2	DAG3	DAG4
500	0.06	0.055	0.09	0.525	0.76	0.145	0.57	0.865	0.31	0.125	0.12	0.33
1000	0.055	0.04	0.185	0.725	0.86	0.225	0.795	0.995	0.255	0.13	0.06	0.3
2000	0.07	0.04	0.32	0.915	0.995	0.42	0.945	1	0.19	0.095	0.075	0.26
4000	0.05	0.02	0.48	1	0.99	0.685	0.98	1	0.18	0.1	0.085	0.3
10000	0.065	0.03	0.805	1	1	0.975	0.995	1	0.14	0.095	0.115	0.355

Violations were introduced in DAG3, which includes unmeasured confounding between A – M and M – Y , and in DAG4, which includes a direct effect of A on Y . See Appendix G.5 for details.

We conducted three sets of simulations. The *first* confirmed that all tests controlled type I error and gained power with increasing sample size under correctly specified models across various variable-type configurations (deferred to Appendix Table 13). The *second* examined model misspecification, highlighting the double-robustness of the DR-CCM test, shown in Table 3. The *third* evaluated the dual and primal tests in continuous-variable settings, with and without Super Learner; while Super Learner mitigated type I error inflation under complex DGPs, it reduced power—likely due to increased estimator variance (deferred to Appendix Table 14).

Simulation 6: Efficiency gain. This simulation evaluated the efficiency of ATE one-step

estimators leveraging the Verma constraint via an anchor variable Z (Section 6.3). We considered two scenarios: (1) binary Z , comparing $\psi_{z^*=1}^+$, $\psi_{z^*=0}^+$, and the optimally weighted estimator $\psi_{\alpha^{\text{opt}}}^+$; and (2) continuous $Z \sim \text{Normal}(1, 1)$, evaluating ψ_p^+ under three choices of $\tilde{p}(Z)$: the true density $p(Z)$, $\text{Normal}(0, 1)$ and $\text{Normal}(10, 1)$. Each setting was replicated 1000 times at sample sizes from 500 to 8000. Full data-generating details are provided in Appendix G.6.

For binary Z , $\psi_{\alpha^{\text{opt}}}^+$ achieved substantially lower variance than either fixed-level estimator, reducing asymptotic variance by nearly half (Appendix Fig. 14). For continuous Z , using $\tilde{p}(Z) = \text{Normal}(10, 1)$ yielded the lowest variance, followed by $p(Z)$ (Appendix Fig. 15). These results illustrate how leveraging the Verma constraint can significantly improve estimator efficiency.

8 Real data application

We applied our front-door estimation framework to two real-world data sets: a longitudinal Finnish cohort examining the effect of early academic performance on future income [Jorma, 2018] (results in Appendix H.2) and an observational study evaluating the impact of mobile stroke unit (MSU) dispatch on post-stroke outcomes in the Berlin prehospital stroke care trial, known as B_PROUD [Ebinger et al., 2017]. We focus on the latter as our primary application below.

The B_PROUD study is a nonrandomized investigation of MSU care conducted in Berlin between February 2017 and May 2019 [Ebinger et al., 2017]. This dataset was previously analyzed by Piccininni et al. [2023] using a front-door approach to estimate the causal effect of MSU dispatch on 3-month functional outcomes. To enable estimation with continuous mediators, their analysis discretized the time from ambulance dispatch to thrombolysis into coarse categories—an approach that, while practical, can lead to information loss and sensitivity to bin definitions. In contrast, our framework accommodates mixed-type mediators without discretization, leveraging flexible machine learning tools to preserve the full resolution of the data.

We applied our method to 768 patients eligible for reperfusion therapy in the B_PROUD cohort, of whom 588 (77%) received MSU care ($A = 1$) and 180 (23%) received conventional emergency services ($A = 0$). The outcome of interest, Y , is the 3-month modified Rankin Scale (mRS) score, an ordinal measure ranging from 0 (no symptoms) to 6 (death). The assumed causal pathway from A to Y is fully mediated through two variables: (i) M_1 , a binary indicator

of thrombolysis receipt, and (ii) M_2 , the time from ambulance dispatch to thrombolysis (set to 0 if thrombolysis was not received). We adjusted for measured baseline covariates: systolic blood pressure (X_1) and stroke severity (X_2).

To handle the ordinal outcome, we constructed binary indicators $Y_k := \mathbb{I}(Y \leq k)$ for $k = 0, \dots, 5$, applied our estimators to each, and recovered the marginal probability mass function $p(Y^a = k)$ by differencing cumulative probabilities. This allowed us to estimate the full distribution of potential outcomes under each treatment level. For comparability with [Piccininni et al. \[2023\]](#), we also replicated their discretization of M_2 using the first quartile and median as cutoffs. Results from this secondary analysis are provided in Appendix [H.1](#).

We estimated the ATE using both the one-step estimator $\psi_{2b}^+(\hat{Q})$ and its TMLE counterpart $\psi_{2b}(\hat{Q}^*)$. To flexibly capture potential nonlinearities and interactions, we used Super Learner with five-fold cross-fitting. The ensemble library included intercept-only models, GLMs, multivariate adaptive regression splines, and random forests.

The one-step estimate of ATE was -0.079 (95% CI: $(-0.468, 0.311)$), while TMLE yielded -0.074 (95% CI: $(-0.464, 0.315)$). Although not statistically significant, both estimates suggest a shift toward improved outcomes with MSU care. To further characterize this effect, we estimated the full potential outcome distributions. Under MSU care, TMLE estimated the following mRS distribution: 0(29%), 1(20%), 2(11%), 3(15%), 4(13%), 5(3%), 6(9%). These estimates are generally consistent with those reported in the original analysis by [Piccininni et al. \[2023\]](#), which found corresponding values of 0(30%), 1(19%), 2(12%), 3(15%), 4(12%), 5(4%), 6(9%).

9 Discussions

While the front-door model provides a powerful framework for causal inference in the presence of treatment-outcome unmeasured confounding, its practical utility depends on both robust estimation strategies and the validity of its identifying assumptions. In this paper, we developed a suite of influence function-based estimators for both the ATE and ATT that accommodate complex, multivariate mediators without relying on parametric assumptions. Our estimators incorporate modern machine learning methods and use sample-splitting to avoid reliance on Donsker conditions, enabling valid inference in flexible settings. Beyond estimation, we also

addressed the testability of key identification assumptions by leveraging a generalized equality constraint involving an anchor variable, which we incorporate into a semiparametric model under the null to both test these assumptions and construct more efficient estimators in this setting.

Despite these advances, several important directions remain. One is to extend our estimation strategies to more complex causal structures, such as hidden variable DAGs represented by acyclic directed mixed graphs. While identification theory in these models is well developed, efficient and flexible estimation remains an open challenge. Expanding one-step and TMLE methods to this setting would improve applicability when mediators only partially explain the treatment effect or unmeasured confounding extends beyond the treatment–outcome link. Another direction is to refine our doubly robust evaluation tools—such as test statistics and confidence intervals—for settings with multiple mediators of mixed types. Finally, extending these methods to longitudinal data with time-varying treatments and mediators would support more realistic analyses where mediation mechanisms and front-door structure evolve over time.

References

- A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of UAI-94*, pages 46–54, 1994.
- M. F. Bellemare, J. R. Bloem, and N. Wexler. The paper of how: Estimating treatment effects using the front-door criterion. Technical report, Working paper, 2019.
- D. Benkeser. *Data-adaptive Estimation in Longitudinal Data Structures with Applications in Vaccine Efficacy Trials*. PhD thesis, 2015.
- D. Benkeser and M. Van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.
- D. Benkeser, M. Carone, M. V. D. Laan, and P. B. Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.
- R. Bhattacharya and R. Nabi. On testability of the front-door model via verma constraints. In *Uncertainty in Artificial Intelligence*, pages 202–212. PMLR, 2022.
- R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23:1–76, 2022.
- G. Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.
- P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.
- I. Díaz, N. S. Hejazi, K. E. Rudolph, and M. J. van Der Laan. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*, 108(3):627–641, 2021.

- M. Ebinger, P. Harmel, C. H. Nolte, U. Grittner, B. Siegerink, and H. J. Audebert. Berlin prehospital or usual delivery of acute stroke care—study protocol. *International Journal of Stroke*, 12(6):653–658, 2017.
- I. R. Fulcher, I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: The generalized front-door criterion. *Journal of the Royal Statistical Society, Series B*, 2019.
- A. N. Glynn and K. Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, 113(523):1040–1049, 2018.
- S. Gruber and M. J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.
- J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.
- T. Hayfield and J. S. Racine. Nonparametric econometrics: The np package. *Journal of statistical software*, 27:1–32, 2008.
- M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.
- K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.
- Y. Huang and M. Valtorta. Pearl’s calculus of interventions is complete. In *Twenty Second Conference On Uncertainty in Artificial Intelligence*, 2006.
- K. Jorma. Life course 1971-2002 [dataset]. version 2.0, 2018. Finnish Social Science Data Archive [distributor]. <http://urn.fi/urn:nbn:fi:fsd:T-FSD2076>.
- T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.
- E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.
- C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.
- J. Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essay des principes. excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472, 1923.
- M. Paschali, Q. Zhao, E. Adeli, and K. M. Pohl. Bridging the gap between deep learning and hypothesis-driven analysis via permutation testing. In *International Workshop on Predictive Intelligence In Medicine*, pages 13–23. Springer, 2022.
- J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.
- M. Piccininni, T. Kurth, H. J. Audebert, and J. L. Rohmann. The effect of mobile stroke unit care on functional outcomes: an application of the front-door formula. *Epidemiology*, 34(5):712–720, 2023.
- T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. 2013.
- T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.
- J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.
- J. M. Robins and L. Wasserman. Estimation of effects of sequential treatments by reparameterizing

- directed acyclic graphs. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 409–420, 1997.
- J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427):846–866, 1994.
- J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.
- P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.
- D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.
- D. O. Scharfstein, R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*, 2021.
- I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.
- I. Shpitser and J. Pearl. Dormant independence. In *Conference on Artificial Intelligence*, volume 23. AAAI Press, 2008.
- M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.
- M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.
- J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002. ISBN 0-262-51129-0.
- A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
- M. J. Van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57, 2014.
- M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.
- M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.
- M. J. van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.
- A. van der Vaart and J. A. Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.
- A. W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.
- T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.
- L. Wen, A. Sarvet, and M. Stensrud. Causal effects of intervening variables in settings with unmeasured confounding. *Journal of Machine Learning Research*, 25(345):1–54, 2024.
- M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013.
- W. Zheng and M. J. Van Der Laan. Asymptotic theory for cross-validated tmle. 2010.
- W. Zheng and M. J. Van Der Laan. Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics*, 8(1):1–40, 2012.

J. Zhou, Z. Zhang, Z. Li, and J. Zhang. Coarsened propensity scores and hybrid estimators for missing data and causal inference. *International Statistical Review*, 83(3):449–471, 2015.

Supplementary Materials

The supplementary materials are structured as follows. Appendix A offers a summary of the notations used throughout the manuscript for ease of reference. After Appendix A, each subsequent appendix provides additional details related to the corresponding section of the paper. Appendix B details the identification proofs of the ATE and ATT under the front-door model and the derivations of the corresponding efficient influence functions. It also includes a brief overview of the geometric views of the front-door statistical model and the breakdown of the tangent space into orthogonal subspaces. Appendices C and D provide additional technical details on the ones-step estimators and TMLE procedures for the ATE and ATT functionals, respectively. These include validations of loss function–submodel pairs, adjustments for binary outcomes, and algorithmic summaries of the TMLE steps. Appendix E presents the proofs underlying inference results for the ATE and ATT estimators, including second-order remainder terms, regularity conditions, and robustness properties. It also includes the formal asymptotic theorems for the ATT estimators. Appendix F provides the technical details for the three testing procedures for front-door assumptions, the construction of more efficient ATE estimators under a semiparametric front-door model, and the Verma constraint, including all relevant identification and estimation proofs. Appendix G presents details of the simulation studies, along with additional simulation results. Appendix H elaborates on the real data application from the main manuscript and contains a second application on assessing the effect of academic performance on future income under the front-door model.

We use the following integration notations interchangeably in the supplementary material:

$$\int(\cdot)dP(x) = \int(\cdot)p(x)dx, \quad \int(\cdot)dP(x, y) = \iint(\cdot)p(x, y)dxdy,$$
 for any random variables X and Y .

A Glossary of terms and notations

To ease navigation of the notations, we provide a comprehensive list in Table 4.

Table 4: Glossary of terms and notations

Symbol	Definition	Symbol	Definition
A, a_0	Treatment, fixed assignment	$\pi(A X)$	propensity score
Y, Y^a	Outcome, potential outcome	$\mu(M, A, X)$	Outcome regression
X	Observed confounders	$f_M(M A, X)$	Mediator density
M	Mediator(s)	$\xi(M, X)$	$\sum_{a \in \{0,1\}} \mu(M, a, X) \pi(a X)$
U	Unmeasured variables	$\eta(A, X)$	$\int \mu(m, A, X) f_M(m a_0, X) dm$
O	Observed data (X, A, M, Y)	$\theta(X)$	$\int \xi(m, X) f_M(m a_0, X) dm$
P	Observed data distribution	$\gamma(X)$	$\mathbb{E}(\xi(M, X) a_0, X) \equiv \theta(X)$
Q	Collection of nuisances	$f_M'(M, A, X)$	$f_M(M a_0, X) / f_M(M A, X)$
$\psi(P)$	Target parameter for ATE ($\equiv \psi(Q)$)	$\lambda(A M, X)$	$p(A M, X)$
$\beta(P)$	Target parameter for ATT ($\equiv \beta(Q)$)	$p_A(A)$	$p(A)$
$\Phi(Q)$	Efficient influence function for $\psi(P)$	$\kappa_a(X)$	$\mathbb{E}(\mu(M, a, X) a_0, X)$
$\Phi_\beta(Q)$	Efficient influence function for $\beta(P)$	$p_{AX}(A, X)$	$p(A, X)$
\hat{Q}	Initial estimate of Q	$\tau(A, X)$	$\mathbb{E}(f_M'(M, A, X) \mu(M, A, X) A, X)$
\hat{Q}^*	TMLE estimate of Q	$H_A(X)$	Clever covariate in treatment model
p_X	Covariates distribution	$H_M(X)$	Clever covariate in mediator model
P_n	Empirical distribution	\mathcal{M}, \mathcal{X}	Domains for variables M, X
L_{Q_j}	Loss function for nuisance $Q_j \in Q$	$\mathcal{M}_{Q_j}, \mathcal{M}_Q$	Model space for nuisance Q_j and Q
$\psi^+(\hat{Q}), \beta^+(\hat{Q})$	One-step estimators	$\psi(\hat{Q}^*), \beta(\hat{Q}^*)$	TMLEs
$\psi_1(\cdot), \beta_1(\cdot)$	Estimators with density estimation	$\psi_{2a}(\cdot), \beta_a(\cdot)$	Estimators with density ratio estimation
$\psi_{2b}(\cdot), \beta_b(\cdot)$	Estimators with Bayes' rule	R_2	Second-order remainder
$\psi_{1,k}^{+,cf}(\hat{Q}^{(-k)})$	k-th fold cross-fitted estimator of ψ_1^+	Z	Anchor variable
q_{primal}	Primal weight	q_{dual}	Dual weight
μ_{primal}^a	MSE risk minimizer with primal weight	μ_{dual}^a	MSE risk minimizer with dual weight
T_n, primal	Primal test statistic	$T_{n, \text{dual}}$	Dual test statistic
$\mu^m(z, x)$	$\mathbb{E}(Y^m Z = z, X = x)$	$T_{n, \text{CCM}}$	Conditional counterfactual mean test statistic
$\psi_{z^*}(Q)$	$\psi(Q)$ at $Z = z^*$	Φ_{z^*}	Efficient influence function for $\psi_{z^*}(Q)$
Λ_α	Class of IFs defined by weight α	α^{opt}	Optimal weight

B Details on causal front-door model

B.1 Nonparametric identification

B.1.1 Identification of $\mathbb{E}(Y^{a_0})$

Given the stated identification assumptions, $p(Y^{a_0} = y)$ can be identified as follows:

$$\begin{aligned}
 p(Y^{a_0} = y) &= \iint p(Y^{a_0} = y, M^{a_0} = m, X = x) dm dx \\
 &= \iint p(Y^m = y | M^{a_0} = m, x) p(M^{a_0} = m | x) p(x) dm dx \\
 &= \iint \left\{ \sum_{a=0}^1 p(Y^m = y, A = a | M^{a_0} = m, x) \right\} p(M^{a_0} = m | x) p(x) dm dx
 \end{aligned}$$

$$\begin{aligned}
&= \iint \left\{ \sum_{a=0}^1 p(Y^m = y \mid A = a, x) p(A = a \mid x) \right\} p(M = m \mid A = a_0, x) p(x) dm dx \\
&= \iint \left\{ \sum_{a=0}^1 p(Y = y \mid M = m, A = a, x) p(A = a \mid x) \right\} p(M = m \mid A = a_0, x) p(x) dm dx,
\end{aligned}$$

where the first equality holds by probability rules, second by factorization rules, and a combination of consistency and no direct effect assumptions, the third holds by probability rules, the fourth holds by factorization rules, consistency, positivity, and conditional ignorability, and the fifth holds by conditional ignorability, consistency, and positivity. Thus, the target parameter $\mathbb{E}(Y^{a_0})$ is identified via the following functional:

$$\psi_{a_0}(P) = \iint \sum_{a=0}^1 y p(y \mid m, a, x) p(a \mid x) p(m \mid a_0, x) p(x) dy dm dx.$$

B.1.2 Identification of $\mathbb{E}(Y^{a_0} \mid A = a_1)$

Similarly, given the stated identification assumptions, $p(Y^{a_0} \mid A = a_1)$ can be identified as follows:

$$\begin{aligned}
&p(Y^{a_0} = y \mid A = a_1) \\
&= \iint p(Y^{a_0} = y, M^{a_0} = m, X = x \mid A = a_1) dm dx \\
&= \iint p(Y^m = y \mid M^{a_0} = m, x, A = a_1) p(M^{a_0} = m \mid x, A = a_1) p(x \mid A = a_1) dm dx \\
&= \iint p(Y^m = y \mid x, A = a_1) p(M^{a_0} = m \mid x, A = a_0) p(x \mid A = a_1) dm dx \\
&= \iint p(Y = y \mid M = m, x, A = a_1) p(M = m \mid x, A = a_0) p(x \mid A = a_1) dm dx,
\end{aligned}$$

where the first and second qualities hold by probability rules, the third holds by ignorability, and the last equality holds by consistency and positivity. Thus, the target parameter $\mathbb{E}(Y^{a_0} \mid A = a_1)$ is identified via the following functional:

$$\beta_{a_0}(P) = \int y p(y \mid m, A = a_1, x) p(m \mid A = a_0, x) p(x \mid A = a_1) dy dm dx.$$

B.1.3 Identification of $\mathbb{E}(Y^{a_1, M^{a_0}} \mid A = a_1)$

In addition to (i) consistency, (ii) conditional ignorability, and (iii) positivity, identification of $\mathbb{E}(Y^{a_1, M^{a_0}} \mid A = a_1)$ requires an additional assumption: (iv) *cross-world independence* stating that $M^{a_0} \perp Y^{a_1, m} \mid A = a_1, X$. Under these assumptions, $\mathbb{E}(Y^{a_1, M^{a_0}} \mid A = a_1)$ is identified as:

$$\begin{aligned} & \mathbb{E}(Y^{a_1, M^{a_0}} \mid A = a_1) \\ &= \iiint y \, p(Y^{a_1, m} = y \mid M^{a_0} = m, A = a_1, x) \, p(M^{a_0} = m \mid A = a_1, x) \, p(x \mid A = a_1) \, dy \, dm \, dx \\ &= \iiint y \, p(Y^{a_1, m} = y \mid A = a_1, x) \, p(M^{a_0} = m \mid A = a_1, x) \, p(x \mid A = a_1) \, dy \, dm \, dx \\ &= \iiint y \, p(Y = y \mid M = m, A = a_1, x) \, p(M^{a_0} = m \mid A = a_1, x) \, p(x \mid A = a_1) \, dy \, dm \, dx \\ &= \iint \mathbb{E}(Y = y \mid M = m, A = a_1, x) \, p(M = m \mid A = a_0, x) \, p(x \mid A = a_1) \, dm \, dx, \end{aligned}$$

where the second equality holds by (iv), and the third and fourth holds by (i) and (ii). Thus, the target parameter $\mathbb{E}(Y^{a_1, M^{a_0}} \mid A = a_1)$ is identifiable via the same functional as $\beta_{a_0}(P)$.

B.2 Alternative interpretations of the front-door functionals

The ATE front-door functional in (1) corresponds to the *population intervention indirect effect* (PIIE) introduced by [Fulcher et al. \[2019\]](#). The PIIE, indexed by a fixed treatment level a_0 , captures the mean difference between Y (the observed outcome) and $Y^{A, M^{a_0}}$ (the counterfactual outcome) under an intervention that shifts the mediator to the value it would have taken had treatment been set to a_0 ; i.e., $\text{PIIE}(a_0) := \mathbb{E}(Y - Y^{A, M^{a_0}})$. Instead of assuming no direct effect of treatment on the outcome—as required by the front-door model—[Fulcher et al. \[2019\]](#) identify the PIIE by replacing this condition with a cross-world independence assumption: $M^{a_0} \perp Y^{a_1, m} \mid A = a_1, X$. Under this alternative assumption, the counterfactual mean $\mathbb{E}(Y^{A, M^{a_0}})$ remains identified by the front-door functional $\psi_{a_0}(P)$ in (1). This connection implies that our proposed estimators, outlined in the next section, retain some meaningful interpretation even when the full mediation assumption fails, thereby broadening their applicability to settings where treatment has both direct and indirect effects.

A closely related interpretation applies to the ATT front-door functional in (2), which corresponds to a PIIE among the treated (PIIE-T) or among the controls (PIIE-C), depending on the

conditioning group, PIIE-T := $\mathbb{E}(Y - Y^{1,M^0} | A = 1)$ and PIIE-C := $\mathbb{E}(Y - Y^{0,M^1} | A = 0)$. The counterfactual parameter $\mathbb{E}(Y^{a_1,M^{a_0}} | A = a_1)$ captures the expected outcome for individuals who received treatment level a_1 , had they retained their treatment assignment but experienced mediator values as if they had received $A = a_0$. This quantity is directly identified by the conditional front-door functional $\beta_{a_0}(P)$ under the same cross-world assumption of [Fulcher et al. \[2019\]](#); see Appendix B.1 for a proof. These interpretations imply that our ATT and ATC estimators also recover subgroup-specific PIIEs, capturing the component of the treatment effect that operates through shifting the values of M under specific interventions within each subpopulation, under alternative assumptions to those required by the standard front-door model.

[Wen et al. \[2024\]](#) provide another interpretation of the front-door functional, viewing it as the *average causal effect on an intervening variable*, defined as $\mathbb{E}(Y^{a_M=1} - Y^{a_M=0})$. Here, A_M represents an intervenable component of the treatment, distinct from the original variable A , which may not correspond to a well-defined or manipulable intervention. In one of their motivating examples, A reflects chronic pain—an inherently non-manipulable construct—that influences a doctor’s perception of the patient’s pain status, captured by A_M . This perceived status in turn affects opioid use (M) and mortality (Y) in their data application. Under identification assumptions, they show that $\mathbb{E}(Y^{a_M=a_0})$ is identified by the same front-door functional $\psi_{a_0}(P)$. This reinforces the relevance of the functional in (1) for policy settings in which direct intervention on A is infeasible, but meaningful action can still be taken on modifiable components such as A_M . Our estimators thus support not only classical mediation analysis, but also modern frameworks that emphasize intervenable causal mechanisms.

These connections substantially broaden the scope of our estimation framework, which remains valid in settings where the effect of A on Y is only partially mediated by M . They also underscore the policy relevance of front-door estimands in scenarios where interventions must target modifiable components of treatment pathways, rather than treatment itself.

B.3 Statistical model and EIF derivations

Let \mathcal{H} denote the *Hilbert space* defined as the space of all mean-zero, square-integrable scalar functions of observed data $O = (X, A, M, Y)$, equipped with the inner product $\mathbb{E}(h_1(O) \times$

$h_2(O)), \forall h_1, h_2 \in \mathcal{H}$. Let \mathcal{M} denote the front-door statistical model, which consists of distributions defined over observed data O . By chain rule of probability, we can write down this joint distribution as $p(o) = p(y | m, a, x) p(m | a, x) p(a | x) p(x)$. Given this factorization, we can write down the joint score as $S(o) = S(y | m, a, x) + S(m | a, x) + S(a | x) + S(x)$.

The *tangent space* of \mathcal{M} , denoted by \mathcal{T} , is defined as the mean-square closure of all linear combinations of scores in corresponding parametric submodels for \mathcal{M} . We can partition \mathcal{T} into a *direct sum* of four orthogonal subspaces, $\mathcal{T} = \mathcal{T}_Y \oplus \mathcal{T}_M \oplus \mathcal{T}_A \oplus \mathcal{T}_X$, defined as follows:

$$\begin{aligned}\mathcal{T}_Y &= \left\{ h_Y(Y, M, A, X) \in \mathcal{H}, \text{ s.t. } \mathbb{E}(h_Y(Y, M, A, X) | M, A, X) = 0 \right\}, \\ \mathcal{T}_M &= \left\{ h_M(M, A, X) \in \mathcal{H}, \text{ s.t. } \mathbb{E}(h_M(M, A, X) | A, X) = 0 \right\}, \\ \mathcal{T}_A &= \left\{ h_A(A, X) \in \mathcal{H}, \text{ s.t. } \mathbb{E}(h_A(A, X) | X) = 0 \right\}, \\ \mathcal{T}_X &= \left\{ h_X(X) \in \mathcal{H}, \text{ s.t. } \mathbb{E}(h_X(X)) = 0 \right\}.\end{aligned}$$

Demonstrating the mutual orthogonality of these tangent spaces is straightforward. For instance, the inner product of any $h_Y(Y, M, A, X) \in \mathcal{T}_Y$ and $h_M(M, A, X) \in \mathcal{T}_M$ is zero, since:

$$\mathbb{E}(h_Y(Y, M, A, X) \times h_M(M, A, X)) = \mathbb{E}(h_M(M, A, X) \times \mathbb{E}(h_Y(Y, M, A, X) | M, A, X)) = 0,$$

which confirms the orthogonality of \mathcal{T}_Y and \mathcal{T}_M . Similar arguments can be applied to prove orthogonality between other pairs of tangent spaces. In the context of the front-door model, where there is no independence restriction among any sets of variables, the tangent space encompasses the entire Hilbert space. Broadly speaking, any statistical model in which \mathcal{T} is equivalent to \mathcal{H} is classified as *nonparametric saturated*.

Any function $h(O)$ within the Hilbert space \mathcal{H} can be *uniquely* decomposed into orthogonal components, expressed as $h = h_Y + h_M + h_A + h_X$. Here, h_V represents the projection of h onto \mathcal{T}_V for each V in the set $\{Y, M, A, X\}$. A prime example of this decomposition is observed in the nonparametric EIF, which is an element in \mathcal{H} . An EIF, say denoted by $\Phi(Q)(O)$, can be broken down into four distinct components, each corresponding to the unique projection of $\Phi(Q)(O)$ onto one of the four mutually orthogonal tangent spaces. The projection $\Phi_Y(Q)(O)$ is specifically shown as a unique projection of $\Phi(Q)(O)$ onto \mathcal{T}_Y . Similar

proofs for $\Phi_M(Q)(O)$, $\Phi_A(Q)(O)$, and $\Phi_X(Q)(O)$ as projections onto \mathcal{T}_M , \mathcal{T}_A , and \mathcal{T}_X , respectively, can be readily formulated. Demonstrating that $\Phi_Y(Q)(O)$ is a projection of $\Phi(O)$ onto \mathcal{T}_Y is equivalent to showing that for any $h_Y(Y, M, A, X) \in \mathcal{T}_Y$, the equation $\mathbb{E}((\Phi(Q)(O) - \Phi_Y(Q)(O))h_Y(Y, M, A, X)) = 0$ holds true. Note that $\Phi(Q)(O) - \Phi_Y(Q)(O)$ is only a function of M, A, X . Thus, via the tower rule, we have: $\mathbb{E}((\Phi(Q)(O) - \Phi_Y(Q)(O))h_Y(Y, M, A, X)) = \mathbb{E}((\Phi(Q)(O) - \Phi_Y(Q)(O))\mathbb{E}(h_Y(Y, M, A, X) | M, A, X)) = 0$.

In the following, we let $o = (x, a, m, y)$ denote realizations of $O = (X, A, M, Y)$.

B.3.1 EIF for the identification functional of $\mathbb{E}(Y^{a_0})$

The EIF for the ID functional of $\mathbb{E}(Y^{a_0})$, denoted by $\psi(Q)$ ($\equiv \psi(P)$), is derived as follows:

$$\begin{aligned}\frac{\partial}{\partial \varepsilon} \psi(P_\varepsilon) \Big|_{\varepsilon=0} &= \frac{\partial}{\partial \varepsilon} \int y dP_\varepsilon(y | m, a, x) dP_\varepsilon(m | a_0, x) dP_\varepsilon(a | x) dP_\varepsilon(x) \Big|_{\varepsilon=0} \\ &= \int y S(y | m, a, x) dP(y | m, a, x) dP(m | a_0, x) dP(a | x) dP(x) \quad (1) \\ &\quad + \int y S(m | a_0, x) dP(y | m, a, x) dP(m | a_0, x) dP(a | x) dP(x) \quad (2) \\ &\quad + \int y S(a, x) dP(y | m, a, x) dP(m | a_0, x) dP(a | x) dP(x). \quad (3)\end{aligned}$$

Given our notations, line (1) simplifies to:

$$\begin{aligned}&\int y S(y | m, a, x) dP(y | m, a, x) dP(m | a_0, x) dP(a | x) dP(x) \\ &= \int f_M^r(m, a, x) [y - \mu(m, a, x)] S(y | m, a, x) dP(y, m, a, x) \\ &= \int f_M^r(m, a, x) [y - \mu(m, a, x)] S(o) dP(o).\end{aligned}$$

Line (2) simplifies to:

$$\begin{aligned}&\int y S(m | a_0, x) dP(y | m, a, x) dP(m | a_0, x) dP(a | x) dP(x) \\ &= \int \sum_a \mu(m, a, x) \pi(a | x) S(m | a_0, x) dP(m | x, a_0) dP(x) \\ &= \int \frac{\mathbb{I}(a = a_0)}{\pi(a | x)} \xi(m, x) S(m | a_0, x) dP(o) \\ &= \int \frac{\mathbb{I}(a = a_0)}{\pi(a | x)} [\xi(m, x) - \theta(x)] S(m | a, x) dP(o)\end{aligned}$$

$$= \int \frac{\mathbb{I}(a = a_0)}{\pi(a | x)} [\xi(m, x) - \theta(x)] S(o) dP(o).$$

Line (3) simplifies to:

$$\begin{aligned} & \int y S(a, x) dP(y | m, a, x) dP(m | a_0, x) dP(a, x) \\ &= \int (\eta(a, x) - \psi) S(a, x) dP(a, x) \\ &= \int (\eta(a, x) - \psi) S(o) dP(o). \end{aligned}$$

Therefore, the EIF for $\psi(Q)$, denoted by $\Phi(Q)(O)$, is:

$$\begin{aligned} \Phi(Q)(O) &= \underbrace{\frac{f_M(M | a_0, X)}{f_M(M | A, X)} \{Y - \mu(M, A, X)\}}_{\Phi_Y(Q)(O)} + \underbrace{\frac{\mathbb{I}(A = a_0)}{\pi(a_0 | X)} \{\xi(M, X) - \theta(X)\}}_{\Phi_M(Q)(O)} \\ &\quad + \underbrace{\eta(A, X) - \theta(X)}_{\Phi_A(Q)(O)} + \underbrace{\theta(X) - \psi(Q)}_{\Phi_X(Q)(O)}. \end{aligned}$$

When A is binary, $\Phi_A(Q)$ can be simplified as:

$$\begin{aligned} \eta(A, X) - \theta(X) &= \sum_{a=0}^1 [\mathbb{I}(A = a) \eta(a, X) - \eta(a, X) \pi(a | X)] \\ &= \sum_{a'=0}^1 \eta(a, X) \{\mathbb{I}(A = a) - \pi(a | X)\} \\ &= \{\eta(1, X) - \eta(0, X)\} \{A - \pi(1 | X)\}. \end{aligned}$$

Similarly, when M is binary, $\Phi_M(Q)$ can be simplified as:

$$\begin{aligned} \frac{\mathbb{I}(A = a_0)}{\pi(a_0 | X)} \{\xi(M, X) - \theta(X)\} &= \frac{\mathbb{I}(A = a_0)}{\pi(a_0 | X)} \sum_{m=0}^1 \{\mathbb{I}(M = m) \xi(m, X) - \xi(m, X) f_M(m | a_0, X)\} \\ &= \frac{\mathbb{I}(A = a_0)}{\pi(a_0 | X)} \sum_{m=0}^1 \xi(m, X) \{\mathbb{I}(M = m) - f_M(m | a_0, X)\} \\ &= \frac{\mathbb{I}(A = a_0)}{\pi(a_0 | X)} \{\xi(1, X) - \xi(0, X)\} \{M - f_M(1 | a_0, X)\}. \end{aligned}$$

B.3.2 EIF for the identification functional of $\mathbb{E}(Y^{a_0} | A = a_1)$

The EIF for the ID functional of $\mathbb{E}(Y^{a_0} | A = a_1)$, denoted by $\beta(P)$ ($\equiv \beta(Q)$), is derived as follows:

$$\begin{aligned}\frac{\partial}{\partial \varepsilon} \beta(P_\varepsilon) \Big|_{\varepsilon=0} &= \frac{\partial}{\partial \varepsilon} \int y dP_\varepsilon(y | m, a_1, x) dP_\varepsilon(m | a_0, x) dP_\varepsilon(x | a_1) \Big|_{\varepsilon=0} \\ &= \int y S(y | m, a_1, x) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) \quad (4)\end{aligned}$$

$$+ \int y S(m | a_0, x) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) \quad (5)$$

$$+ \int y S(x | a_1) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) . \quad (6)$$

Given our notations, line (4) simplifies to:

$$\begin{aligned}&\int y S(y | m, a_1, x) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) \\ &= \int \frac{\mathbb{I}(a = a_1)}{p_A(a_1)} \frac{f_M(m | a_0, x)}{f_M(m | a_1, x)} [y - \mu(m, a_1, x)] S(y | m, a, x) dP(y, m, a, x) \\ &= \int \frac{\mathbb{I}(a = a_1)}{p_A(a_1)} \frac{f_M(m | a_0, x)}{f_M(m | a, x)} [y - \mu(m, a_1, x)] S(o) dP(o) .\end{aligned}$$

Line (5) simplifies to:

$$\begin{aligned}&\int y S(m | a_0, x) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) \\ &= \int \frac{\mathbb{I}(a = a_0)}{p_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} [\mu(m, a_1, x) - \kappa_{a_1}(x)] S(m | a, x) dP(m, a, x) \\ &= \int \frac{\mathbb{I}(a = a_0)}{p_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} [\mu(m, a_1, x) - \kappa_{a_1}(x)] S(o) dP(o) .\end{aligned}$$

Line (6) simplifies to:

$$\begin{aligned}&\int y S(x | a_1) dP(y | m, a_1, x) dP(m | a_0, x) dP(x | a_1) \\ &= \int \frac{\mathbb{I}(a = a_1)}{p_A(a_1)} [\kappa_{a_1}(x) - \beta] S(x | a) dP(x, a) \\ &= \int \frac{\mathbb{I}(a = a_1)}{p_A(a_1)} [\kappa_{a_1}(x) - \beta] S(o) dP(o) .\end{aligned}$$

Therefore, the EIF for $\beta(Q)$, denoted by $\Phi_\beta(Q)(O)$ is given by:

$$\begin{aligned}\Phi_\beta(Q)(O) &= \underbrace{\frac{\mathbb{I}(a = a_1)}{p_A(a_1)} \frac{f_M(m | a_0, x)}{f_M(m | a, x)} [y - \mu(m, a_1, x)]}_{\Phi_{\beta,Y}(Q)(O)} + \underbrace{\frac{\mathbb{I}(a = a_0)}{p_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} [\mu(m, a_1, x) - \kappa_{a_1}(x)]}_{\Phi_{\beta,M}(Q)(O)} \\ &\quad + \underbrace{\frac{\mathbb{I}(a = a_1)}{p_A(a_1)} [\kappa_{a_1}(x) - \beta]}_{\Phi_{\beta,AX}(Q)(O)}.\end{aligned}$$

B.4 Overview of one-step corrected plug-in estimation

The stochastic behavior of a plug-in estimator $\psi(\hat{Q})$ can be studied using a linear expansion of the parameter. Given an P -integrable function f of the observed data O , let $Pf := \int f(o)p(o)do$ and $P_n f := \frac{1}{n} \sum_{i=1}^n f(O_i)$. A linear expansion of $\psi(\hat{Q})$ yields $\psi(\hat{Q}) = \psi(Q) - P\Phi(\hat{Q}) + R_2(\hat{Q}, Q)$, where Φ is a gradient of ψ satisfying $P\Phi(Q) = 0$, and $R_2(\hat{Q}, Q)$ denotes a second-order remainder term. While multiple gradients may satisfy the expansion in general, the tangent space of our model is saturated such that there is only a single, unique gradient. This gradient is also known as the efficient influence function (EIF) due to its foundational link to the theory of regular, asymptotically linear estimators [Bickel et al., 1993].

To better characterize the stochastic behavior of $\psi(\hat{Q})$, we rewrite its linear expansion as

$$\psi(\hat{Q}) - \psi(Q) = P_n \Phi(Q) - P_n \Phi(\hat{Q}) + (P_n - P)\{\Phi(\hat{Q}) - \Phi(Q)\} + R_2(\hat{Q}, Q). \quad (42)$$

The *first* term in (42) is a sample average of mean-zero i.i.d. terms and thus enjoys standard root- n asymptotic behavior. The *third* term is an empirical process term, which can be shown to be $o_p(n^{-1/2})$ if $\Phi(\hat{Q}) - \Phi(Q)$ falls in a P -Donsker class with probability tending to 1 and $P\{\Phi(\hat{Q}) - \Phi(Q)\}^2 = o_p(1)$ [van der Vaart and Wellner, 2023]. In Section 5, we use sample-splitting procedure to assure that the third term is $o_p(n^{-1/2})$, even if Donsker conditions are not met [Kennedy, 2022, Chernozhukov et al., 2017]. The *fourth* term is the second-order remainder, which can generally be bounded by the convergence rates of respective components of \hat{Q} to their true counterparts. To precisely bound the second-order remainder, we must consider its explicit form. We characterize this remainder in Section 5. For the time being, it suffices to state that if the rates of convergence of nuisance estimators are sufficiently fast, then we generally expect $R_2(\hat{Q}, Q) = o_p(n^{-1/2})$. Finally, the *second* term in (42) is the first-order bias of the

plug-in estimator. When flexible nuisance estimation strategies are used (e.g., based on machine learning), this term may not have standard root- n asymptotic behavior. This motivates the one-step corrected plug-in estimator, denoted by $\psi_1^+(\hat{Q})$, to be $\psi(\hat{Q}) + P_n\Phi(\hat{Q})$.

B.5 Overview of the TMLE framework

Given a plug-in estimator $\psi(\hat{Q})$ of the parameter of interest $\psi(Q)$, the core idea of a TMLE procedure is to find a replacement for \hat{Q} , say \hat{Q}^* , such that the following two aims hold:

- (I) \hat{Q}^* is at least as good an estimate of Q as \hat{Q} , w.r.t. a valid measure of empirical risk,
- (II) $P_n\Phi(\hat{Q}^*) = o_p(n^{-1/2})$, so that the first-order bias of $\psi(\hat{Q}^*)$ would be negligible.

Consider the general setting where $\psi(Q)$ is the parameter of interest and Q is parameterized as (Q_1, Q_2, \dots, Q_J) , i.e., there are J key nuisance parameters needed to evaluate ψ and its EIF. We assume Q belongs in a functional space \mathcal{Q} , defined as $\mathcal{M}_{Q_1} \times \mathcal{M}_{Q_2} \times \dots \times \mathcal{M}_{Q_J}$, i.e., the Cartesian product of the functional spaces of each nuisance functional, denoted by \mathcal{M}_Q . Suppose also that the EIF can be written as $\Phi = \sum_{j=1}^J \Phi_j$, where Φ_j is the component of Φ that belongs to the tangent space associated with Q_j . For example, for $\psi(Q)$ in (1), we can set $Q = (\mu, f_M, \pi, p_X)$, and according to the EIF in (4) $\Phi_1 = \Phi_Y, \Phi_2 = \Phi_M, \Phi_3 = \Phi_A, \Phi_4 = \Phi_X$.

To achieve both aims (I)-(II), the TMLE procedure comprises two main steps: the *initialization* step, where the initial estimate \hat{Q} is obtained, and the subsequent *targeting* step, where \hat{Q} is updated to a new estimate \hat{Q}^* . In the *initialization* step, we obtain an initial estimate of Q based on a collection of estimates for each nuisance parameter individually, $\hat{Q} = (\hat{Q}_1, \dots, \hat{Q}_J)$. In the *targeting* step, we require (i) a *submodel* and (ii) a *loss function* for each component Q_j of Q . For requirement (i), with an estimate \tilde{Q} of Q , we define a submodel $\{\hat{Q}_j(\varepsilon_j; \tilde{Q}_{-j}), \varepsilon_j \in \mathbb{R}\}$ within \mathcal{M}_{Q_j} . This submodel is indexed by a univariate real-valued parameter ε_j and may also depend on \tilde{Q}_{-j} (the components of \tilde{Q} excluding component j) or a subset of \tilde{Q}_{-j} (including the possibility of an empty subset). For requirement (ii), with a given $\tilde{Q} \in \mathcal{Q}$, we denote a loss function for \tilde{Q}_j by $L(\tilde{Q}_j; \tilde{Q}_{-j}) : \mathcal{O} \rightarrow \mathbb{R}$, where \mathcal{O} denotes the state space of the observed data. Note that the loss function for \tilde{Q}_j can also be indexed by \tilde{Q}_{-j} , or possibly by a subset of \tilde{Q}_{-j} , which may sometimes be an empty set. The submodel and loss function must be chosen to satisfy:

$$(\mathbf{C1}) \quad \hat{Q}_j(0; \tilde{Q}_{-j}) = \hat{Q}_j,$$

$$(\mathbf{C2}) \quad Q_j = \operatorname{argmin}_{\tilde{Q}_j \in \mathcal{M}_{Q_j}} \int L(\tilde{Q}_j; Q_{-j})(o) p(o) do,$$

$$(\mathbf{C3}) \quad \frac{\partial}{\partial \varepsilon_j} L(\hat{Q}_j(\varepsilon_j; \hat{Q}_{-j}); \hat{Q}_{-j})|_{\varepsilon_j=0} = \Phi_j(\hat{Q}).$$

(C1) implies that the submodel aligns with the given estimate \hat{Q}_j at $\varepsilon_j = 0$; (C2) indicates that the expectation of the loss function under the true distribution P is minimized at Q_j ; and (C3) ensures that the evaluation of the derivative of the loss function with respect to ε_j at 0 is equivalent to evaluation of the corresponding component of the EIF at \hat{Q} .

Given appropriate choices of submodels and loss functions, we proceed to update \hat{Q} via an iterative risk minimization process. Given current estimates at iteration t , say $\hat{Q}^{(t)}$, we update $\hat{Q}_j^{(t)}$ via empirical risk minimization along the selected submodel using the selected loss function. That is, we define $\hat{\varepsilon}_j = \operatorname{argmin}_{\varepsilon_j \in \mathbb{R}} P_n L(\hat{Q}_j(\varepsilon_j; \hat{Q}_{-j}^{(t)}); \hat{Q}_{-j}^{(t)})$ to be the value of ε_j that minimizes empirical risk given current estimates $\hat{Q}_{-j}^{(t)}$. Condition (C2) suggests that the updated estimate $\hat{Q}_j^{(t+1)} = \hat{Q}_j(\hat{\varepsilon}_j; \hat{Q}_{-j}^{(t)})$ should satisfy (I), as $\hat{Q}_j^{(t+1)}$ will have lower empirical risk than $\hat{Q}_j^{(t)}$. This process is repeated for each of the J components of Q resulting in an updated estimate $\hat{Q}^{(t+1)}$. Condition (C3) suggests that if during this updating process we have found that $\hat{\varepsilon}_j \approx 0$ for each j , then we might expect $P_n \Phi_j(\hat{Q}^{(t+1)}) \approx 0$ for each j and thus (II) may be satisfied. If after iteration t , we find that (II) is not approximately satisfied, we would repeat the updating process. The process is repeated until $P_n \Phi(\hat{Q}^{(t)}) < C_n$, where $C_n = o_p(n^{-1/2})$, e.g., $C_n = \{n^{1/2} \log(n)\}^{-1}$. The final estimate of Q is denoted as \hat{Q}^* and the TMLE is defined as the plug-in estimator $\psi(\hat{Q}^*)$.

We derive TMLEs for all three representations of the ATE and ATT front-door functionals. These estimators differ in both stages of the TMLE procedure: (i) they use different parameterizations of the nuisance functions comprising Q , requiring distinct estimation strategies, and (ii) they employ different techniques to achieve (II), the TMLE approximate-equation-solving property where $P_n \Phi(\hat{Q}^*) = o_p(n^{-1/2})$. Further methodological details are provided in Sections 3 and 4. For a general overview of the TMLE framework, see [van der Laan et al. \[2011\]](#).

C Details on estimators for the ATE front-door functional

C.1 Validity of loss function and submodel combinations

We establish the validity of the loss function and submodel combinations (discussed in Appendix B.5) for the binary mediator case, as detailed in Algorithm 1 (Appendix C.6) and discussed in Section 3.1. Similar proofs for the remaining TMLE procedures follow analogously.

Since the proof for the f_M update closely mirrors that for the propensity score π , we focus here on verifying conditions (C1)–(C3) for the updates to π and μ .

Loss function and submodel combination used for updating π :

$$\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(t)})(1 | X) = \text{expit} \left\{ \text{logit}\{\hat{\pi}^{(t)}(1 | X)\} + \varepsilon_A \left\{ \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X) \right\} \right\}, \varepsilon_A \in \mathbb{R},$$

$$L_A(\tilde{\pi})(O) = -\log \tilde{\pi}(A | X).$$

Proof of (C1): $\hat{\pi}(\varepsilon_A = 0; \hat{\mu}^{(0)}, \hat{f}_m^{(t)})(1 | X) = \text{expit} \left\{ \text{logit}\{\hat{\pi}^{(t)}(1 | X)\} \right\} = \hat{\pi}^{(t)}(1 | X)$.

Proof of (C2): $\mathbb{E}(L_A(\tilde{\pi})(O)) = \mathbb{E}(-\log \tilde{\pi}(A | X)) = \int \{-\sum_a \pi(a | x) \log \tilde{\pi}(a | x)\} dP(x)$ is minimized if $-\sum_a \pi(a | x) \log \tilde{\pi}(a | x)$ is minimized for any $x \in \mathcal{X}$. Since

$$\begin{aligned} -\sum_a \pi(a | x) \log \tilde{\pi}(a | x) &= -\sum_a \pi(a | x) \log \left(\frac{\tilde{\pi}(a | x)}{\pi(a | x)} \times \pi(a | x) \right) \\ &= -\sum_a \pi(a | x) \log \frac{\tilde{\pi}(a | x)}{\pi(a | x)} - \sum_a \pi(a | x) \log \pi(a | x), \end{aligned}$$

we only need to focus on the minimization of $-\sum_a \pi(a | x) \log \frac{\tilde{\pi}(a | x)}{\pi(a | x)}$, which corresponds to the Kullback-Leibler (KL) divergence from $\pi(a | x)$ to $\tilde{\pi}(a | x)$, denoted by $D_{\text{KL}}(\pi || \tilde{\pi})$. This KL-divergence is minimized if $\tilde{\pi}(A | X = x) = \pi(A | X = x)$, for all $x \in \mathcal{X}$.

Proof of (C3):

$$\begin{aligned} \frac{\partial}{\partial \varepsilon_A} L_A(\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(0)})) \Big|_{\varepsilon_A=0} \\ = -\frac{\partial}{\partial \varepsilon_A} \left\{ A \log \hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(0)}) + (1 - A) \log \left\{ 1 - \hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(t)}) \right\} \right\} \Big|_{\varepsilon_A=0} \end{aligned}$$

$$\begin{aligned}
&= - \left\{ A \frac{\partial}{\partial \varepsilon_A} \hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(0)}) + (1 - A) \frac{-\frac{\partial}{\partial \varepsilon} \hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(0)})}{1 - \hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_m^{(0)})} \right\} \Big|_{\varepsilon_A=0} \\
&= \left\{ \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X) \right\} \left\{ \hat{\pi}^{(t)}(1 | X) - A \right\} \propto \Phi_A(\hat{Q}^{(t)}).
\end{aligned}$$

Loss function and submodel combination used for updating μ :

$$\begin{aligned}
\hat{\mu}(\varepsilon_Y)(M, A, X) &= \hat{\mu}^{(t)}(M, A, X) + \varepsilon_Y, \quad \varepsilon_Y \in \mathbb{R}, \\
L_Y(\tilde{\mu}; \hat{f}_M^{(t)})(O) &= \frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} \{Y - \tilde{\mu}(M, A, X)\}^2.
\end{aligned}$$

Proof of (C1): $\hat{\mu}(\varepsilon_Y = 0)(M, A, X) = \hat{\mu}^{(t)}(M, A, X)$.

Proof of (C2):

$$\begin{aligned}
&\mathbb{E}(L_Y(\tilde{\mu}; \hat{f}_M^{(t)})(O)) \\
&= \mathbb{E} \left(\frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} \{Y - \tilde{\mu}(M, A, X)\}^2 \right) \\
&= \mathbb{E} \left(\frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} \{Y - \mu(M, A, X)\}^2 + \frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} \{\mu(M, A, X) - \tilde{\mu}(M, A, X)\}^2 \right),
\end{aligned}$$

which is minimized when $\tilde{\mu}(M, A, X) = \mu(M, A, X)$.

Proof of (C3):

$$\frac{\partial}{\partial \varepsilon} L_Y(\hat{\mu}(\varepsilon_Y; \hat{f}_M^{(t)})) \Big|_{\varepsilon=0} = 2 \frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} (Y - \hat{\mu}^{(t)}(M, A, X)) \propto \Phi_Y(\hat{Q}^{(t)}).$$

C.2 TMLE considerations for binary outcome

For binary outcomes, the TMLE procedure for computing $\psi_1(\hat{Q}^*)$ —originally described in Section 3.1 for continuous outcomes—requires the following modifications.

We adopt a new loss function and submodel for updating $\hat{\mu}$:

$$\hat{\mu}(\varepsilon_Y; \hat{f}_M^{(t)})(M, A, X) = \text{expit} \left\{ \text{logit } \hat{\mu}^{(t)}(M, A, X) + \varepsilon_Y \frac{\hat{f}_M^{(t)}(M | a_0, X)}{\hat{f}_M^{(t)}(M | A, X)} \right\}, \quad \varepsilon_Y \in \mathbb{R}, \quad (43)$$

$$L_Y(\tilde{\mu}) = -\log \tilde{\mu}(M, A, X).$$

Due to the nonlinear nature of the parametric submodel in (43) with respect to ε_Y , computations of $\hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X)$ and $\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)$ would depend on updated estimate of $\hat{\mu}^{(t)}$. Therefore, unlike the continuous outcome case, the dependence of submodels $\hat{\pi}(\varepsilon_A; \hat{\mu}^{(t)}, \hat{f}_m^{(t)})$ and $\hat{f}_M(\varepsilon_M; \hat{\mu}^{(t)}, \hat{\pi}^{(t)})$ on $\hat{\mu}^{(t)}$ would be through the updated estimate $\hat{\mu}^{(t)}$. This implies that once the estimate of μ is updated, the estimates for f_M and π must be updated accordingly. Given $\hat{Q}^{(t)} = (\hat{\mu}^{(t)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$, we modify Step 2 of the continuous outcome case, discussed in Section 3.1, as follows.

Step 2a: Update $\hat{\pi}$, by following the exact same procedure as the one discussed in Section 3.1, modula the fact that $\hat{\mu}$ is replaced with $\hat{\mu}^{(t)}$. After performing the empirical risk minimization and obtaining $\hat{\varepsilon}_A$, we update $\hat{\pi}^{(t+1)} = \pi(\hat{\varepsilon}_A; \hat{\mu}^{(t)}, \hat{f}_M^{(t)})$ and define $\hat{Q}^{(\text{temp}_1)} = (\hat{\mu}^{(t)}, \hat{\pi}^{(t+1)}, \hat{f}_M^{(t)}, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi_A(\hat{Q}^{(\text{temp}_1)}) = o_p(n^{-1/2})$.

Step 2b: Update \hat{f}_M , by following the exact same procedure as the one discussed in Section 3.1, modula the fact that $\hat{\mu}$ is replaced with $\hat{\mu}^{(t)}$. After performing the empirical risk minimization and obtaining $\hat{\varepsilon}_M$, we update $\hat{f}_M^{(t+1)} = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu}^{(t)}, \hat{\pi}^{(t+1)})$ and define $\hat{Q}^{(\text{temp}_2)} = (\hat{\mu}^{(t)}, \hat{\pi}^{(t+1)}, \hat{f}_M^{(t+1)}, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi_M(\hat{Q}^{(\text{temp}_2)}) = o_p(n^{-1/2})$.

Step 2c: Update $\hat{\mu}$, by performing an empirical risk minimization to find

$$\hat{\varepsilon}_Y = \operatorname{argmin}_{\varepsilon_Y \in \mathbb{R}} P_n L_Y(\hat{\mu}(\varepsilon_Y; \hat{f}_M^{(t+1)})). \quad (44)$$

This corresponds to fitting a logistic regression without an intercept term:

$$Y \sim \text{offset}(\text{logit } \hat{\mu}^{(t)}) + \hat{H}_Y^{(t)}(M, A, X), \quad \text{where } \hat{H}_Y^{(t)}(M, A, X) := \frac{\hat{f}_M^{(t+1)}(M | a_0, X)}{\hat{f}_M^{(t+1)}(M | A, X)}.$$

The coefficient of $\hat{H}_Y^{(t)}(M, A, X)$ corresponds to $\hat{\varepsilon}_Y$ as a solution to (44). We update $\hat{\mu}^{(t+1)} = \hat{\mu}(\hat{\varepsilon}_Y; \hat{f}_M^{(t+1)})$, and define $\hat{Q}^{(t+1)} = (\hat{\mu}^{(t+1)}, \hat{\pi}^{(t+1)}, \hat{f}_M^{(t+1)}, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi_Y(\hat{Q}^{(t+1)}) = o_p(n^{-1/2})$. We increment t and repeat Step 2 until convergence.

Assume convergence at iteration t^* . Let $\hat{\pi}^* = \hat{\pi}^{(t^*)}$, $\hat{f}_M^* = \hat{f}_M^{(t^*)}$, $\hat{\mu}^* = \hat{\mu}^{(t^*)}$, and define $\hat{Q}^* = (\hat{\mu}^*, \hat{\pi}^*, \hat{f}_M^*, \hat{p}_X)$. The TMLE plug-in is then given by $\psi_1(\hat{Q}^*)$, as described in (12).

The TMLE procedure for computing $\psi_2(\hat{Q}^*)$ —originally described in Section 3.2 for continuous

outcomes—remains largely unchanged for binary outcomes, with the submodel–loss function pair in (43) used for updating $\hat{\mu}$.

C.3 An alternative submodel for targeting $\hat{\mu}$ under continuous Y

The TMLEs proposed for continuous Y in the main manuscript rely on linear parametric submodels for targeting $\hat{\mu}$. However, such models may exhibit instability in sparse data settings with low Fisher information, as demonstrated in simulations by Zhou et al. [2015]. To address this issue, Gruber and van der Laan [2010] showed that TMLEs using parametric submodels constrained to remain within the semiparametric model of the observed data distribution tend to be more robust than those based on linear submodels. Motivated by this, we introduce an alternative TMLE that employs a nonlinear submodel for targeting the outcome regression μ , when Y is continuous. We outline the key ideas for this construction below, noting that the procedure closely parallels that described in Appendix Section C.2.

Let a and b denote the minimum and maximum observed values of Y , respectively. To enable the use of nonlinear submodels designed for binary outcomes, we rescale Y to the unit interval by defining $Y^* = (Y - a)/(b - a)$, so that $Y^* \in [0, 1]$. Targeting is then performed using Y^* in place of Y , applying the nonlinear parametric submodels defined in (43). All remaining steps of the TMLE procedure follow exactly as described in Appendix C.2. Finally, to return to the original scale, we multiply the point estimate by $(b - a)$ and add a , and rescale the estimated EIF by multiplying it by $(b - a)$.

C.4 Valid submodels for conditional density of a continuous mediator

To ensure that the submodel in (14) is a valid submodel of \mathcal{M}_{f_M} , the range of ε_M must be restricted so that the submodel defines a valid probability density function; that is, $\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})(M | a_0, X) \geq 0$ for all $\varepsilon_M \in (-\delta, \delta)$.

Recall that $\hat{\xi}^{(t)}(M, X) = \sum_{a=0}^1 \hat{\mu}^{(0)}(M, a, X) \hat{\pi}^{(t)}(a | X)$ and $\hat{\theta}^{(t)}(X) = \int \hat{\xi}^{(t)}(m, X) \hat{f}_M^{(t)}(m | a_0, X) dm$.

Let $S_{\text{pos}}^{(t)}$ denote the set of indices for observations with

$$\frac{\hat{\xi}^{(t)}(M_i, X_i) - \hat{\theta}^{(t)}(X_i)}{\hat{\pi}^{(t)}(a_0 | X_i)} > 0.$$

For $i \in S_{\text{pos}}^{(t)}$, $\hat{f}_M(\varepsilon_M, \hat{\mathbf{Q}}^{(t)})(M | a_0, X) \geq 0$ implies that $\varepsilon_M \geq L_i^{(t)}$, where

$$L_i^{(t)} := -\frac{\hat{\pi}^{(t)}(a_0 | X_i)}{\hat{\xi}^{(t)}(M_i, X_i) - \hat{\theta}^{(t)}(X_i)}.$$

Similarly, define $S_{\text{neg}}^{(t)}$ to be the set of indices for observations with

$$\frac{\hat{\xi}^{(t)}(M_i, X_i) - \hat{\theta}^{(t)}(X_i)}{\hat{\pi}^{(t)}(a_0 | X_i)} < 0.$$

For $i \in S_{\text{neg}}^{(t)}$, $\hat{f}_M(\varepsilon_M, \hat{\mathbf{Q}}^{(t)})(M | a_0, X) \geq 0$ implies that $\varepsilon_M \leq R_i^{(t)}$, where

$$R_i^{(t)} := -\frac{\hat{\pi}^{(t)}(a_0 | X_i)}{\hat{\xi}^{(t)}(M_i, X_i) - \hat{\theta}^{(t)}(X_i)}.$$

Let $L^{(t)} = \text{argmax}_{i \in S_{\text{pos}}^{(t)}} L_i^{(t)}$ and $R^{(t)} = \text{argmin}_{i \in S_{\text{neg}}^{(t)}} R_i^{(t)}$. For the given dataset, (L, R) constitutes a valid domain for ε_M . For any $\varepsilon_M \in (L, R)$, we have $\hat{f}_M(\varepsilon_M; \hat{\mu}, \hat{\pi}^{(t)})(M | a_0, X) \geq 0$. Any selection of δ ensuring $(-\delta, \delta) \subseteq (L, R)$ would be applicable for carrying out the TMLE procedure. Note that the valid domain for ε_M changes over iteration alongside the iterative updates of estimates for f_M and π . Consequently, the choice of δ should be relatively small to guarantee the submodel defined in (14) is a valid submodel over all iterations.

Alternatively, we may use the following submodel where ε_M can span the entire real line,

$$\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})(M | a_0, X) = \frac{\hat{f}_M^{(t)}(M | a_0, X) \exp \left[\frac{\varepsilon_M}{\hat{\pi}^{(t)}(a_0 | X)} (\hat{\xi}^{(t)}(M, X) - \hat{\theta}^{(t)}(X)) \right]}{\iint \hat{f}_M^{(t)}(m | a_0, x) \exp \left[\frac{\varepsilon_M}{\hat{\pi}^{(t)}(a_0 | x)} (\hat{\xi}^{(t)}(m, x) - \hat{\theta}^{(t)}(x)) \right] dm dx}. \quad (45)$$

This alternative submodel increases computational complexity, as the denominator must be numerically approximated at each iteration.

C.5 TMLEs that avoid mediator density estimation

Given initial estimates \hat{Q} , a TMLE version of $\psi_2^+(\hat{Q})$ can be formulated as follows.

Step 1: Define loss functions and submodels through $\hat{\mu}, \hat{\pi}, \hat{\gamma}$. Given $\hat{Q} \in \mathcal{Q}, \varepsilon_Y, \varepsilon_A, \varepsilon_\gamma \in \mathbb{R}$, define

$$\begin{aligned}\hat{\mu}(\varepsilon_Y)(M, A, X) &= \hat{\mu}(M, A, X) + \varepsilon_Y, \\ \hat{\pi}(\varepsilon_A; \hat{\kappa})(1 | X) &= \text{expit} \left[\text{logit} \{ \hat{\pi}(1 | X) \} + \varepsilon_A \{ \hat{\kappa}_1(X) - \hat{\kappa}_0(X) \} \right], \\ \hat{\gamma}(\varepsilon_\gamma)(X) &= \hat{\gamma}(X) + \varepsilon_\gamma.\end{aligned}\tag{46}$$

For a given $\tilde{\mu} \in \mathcal{M}_\mu$, $\tilde{\pi} \in \mathcal{M}_\pi$, and $\tilde{\gamma} \in \mathcal{M}_\gamma$, define the following loss functions:

$$\begin{aligned}L_Y(\tilde{\mu}; \hat{f}_M^r)(O) &= \hat{f}_M^r(M, A, X) \{Y - \tilde{\mu}(M, A, X)\}^2, \quad L_A(\tilde{\pi})(O) = -\log \tilde{\pi}(A | X), \\ L_\gamma(\tilde{\gamma}; \hat{\pi}, \hat{\xi})(O) &= \frac{\mathbb{I}(A = a_0)}{\hat{\pi}(a_0 | X)} \left(\hat{\xi}(M, X) - \tilde{\gamma}(X) \right)^2.\end{aligned}\tag{47}$$

See Appendix C.1 for a proof of validity of these submodel–loss function pairs under (C1)–(C3).

Note that the submodel $\hat{\pi}(\varepsilon_A; \hat{\kappa})$ is indexed by $\hat{\kappa}$, which in turn depends on $\hat{\mu}$. However, this submodel remains invariant to updates of $\hat{\mu}$ due to the linearity of the μ submodel in ε_Y , which makes $\hat{\kappa}_1(X) - \hat{\kappa}_0(X)$ effectively fixed by the initial $\hat{\mu}$. Moreover, since the submodels and loss functions for $\hat{\pi}$ and $\hat{\mu}$ are independent of each other's updates, their targeting steps can be performed simultaneously in a single step. In contrast, the submodel and loss function for $\hat{\gamma}$ depend on the targeted versions of $\hat{\pi}$ and $\hat{\mu}$. Thus, targeting $\hat{\gamma}$ must follow the updates of $\hat{\pi}$ and $\hat{\mu}$, using $\hat{\xi}$ and $\hat{\gamma}$ computed from those updated estimates. This sequencing ensures that $\hat{\gamma}$ is targeted using the most recent nuisance values.

Step 2: Perform empirical risk minimizations using submodels and loss functions for μ and π .

Step 2a: Update an estimate of μ by performing an empirical risk minimization to find $\hat{\varepsilon}_Y = \arg\min_{\varepsilon_Y \in \mathbb{R}} P_n L_Y(\hat{\mu}(\varepsilon_Y); \hat{f}_M^r)$. This minimization problem can be solved by fitting $Y \sim \text{offset}(\hat{\mu}) + 1$ with weight $\hat{f}_M^r(M, A, X)$. The intercept coefficient corresponds to $\hat{\varepsilon}_Y$ as the minimizer of the empirical risk. Define $\hat{\mu}^* = \hat{\mu}(\hat{\varepsilon}_Y; \hat{f}_M^r)$ and let $\hat{Q}^{(1)} = (\hat{\mu}^*, \hat{\gamma}, \hat{f}_M^r, \hat{\kappa}, \hat{\pi}, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi_Y(\hat{Q}^{(1)}) = 0$.

Step 2b: Update an estimate of π by performing an empirical risk minimization to find $\hat{\varepsilon}_A = \arg\min_{\varepsilon_A \in \mathbb{R}} P_n L_A(\hat{\pi}(\varepsilon_A; \hat{\kappa}))$. The solution is obtained by fitting the following logistic regression

without an intercept term:

$$A \sim \text{offset}(\text{logit } \hat{\pi}(1 | X)) + \hat{H}_A(X), \text{ where } \hat{H}_A(X) = \hat{\kappa}_1(X) - \hat{\kappa}_0(X).$$

The coefficient in front of the clever covariate $\hat{H}_A(X)$ corresponds to $\hat{\varepsilon}_A$ as the minimizer of the empirical risk. Define $\hat{\pi}^* = \pi(\hat{\varepsilon}_A; \hat{\mu})$ and let $\hat{Q}^{(2)} = (\hat{\mu}^*, \hat{\gamma}, \hat{f}_M^r, \hat{\kappa}, \hat{\pi}^*, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi_A(\hat{Q}^{(2)}) = 0$. Compute $\hat{\gamma}(X)$ by fitting the following linear regression using only data points where $A_i = a_0$ and making prediction using all the data points of X :

$$\hat{\xi}^*(M, X) \sim X, \text{ where } \hat{\xi}^*(M, X) = \sum_{a=0}^1 \hat{\mu}^*(M, a, X) \hat{\pi}^*(a | X).$$

Step 3: Perform one-step risk minimization using pre-defined submodel and loss function for γ .

Update γ by performing an empirical risk minimization to find

$$\hat{\varepsilon}_\gamma = \underset{\varepsilon_\gamma \in \mathbb{R}}{\operatorname{argmin}} P_n L_\gamma \left(\hat{\gamma}(\varepsilon_\gamma); \hat{\pi}^*, \hat{\xi}^* \right). \quad (48)$$

The solution can be obtained by fitting $\hat{\xi}^* \sim \text{offset}(\hat{\gamma}) + 1$ with weight $\frac{\mathbb{I}(A=a_0)}{\hat{\pi}^*(a_0 | X)}$. The intercept coefficient corresponds to $\hat{\varepsilon}_\gamma$ as a solution to the optimization problem in (48). Define $\hat{\gamma}^* = \hat{\gamma}(\hat{\varepsilon}_\gamma)$ and let $\hat{Q}^* = (\hat{\mu}^*, \hat{\gamma}^*, \hat{f}_M^r, \hat{\kappa}, \hat{\pi}^*, \hat{p}_X)$. Condition (C3) implies that $P_n \Phi(\hat{Q}^*) = 0$.

Step 4: Evaluate the plug-in estimator in (15) based on updated estimate $\hat{\gamma}^$,*

$$\psi_2(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}^*(X_i). \quad (49)$$

Remark. One can also adopt an alternative sequential regression for $\theta(X)$, redefined as $\sum_{a=0}^1 \eta(a, X) \pi(a | X)$, with $\eta(a, X) = a \kappa_1(X) + (1-a) \kappa_0(X)$. This reverses the integration order in (1), marginalizing over M first to derive $\eta(A, X)$, rather than over A to obtain $\xi(M, X)$. The resulting plug-in estimator, $\psi_3(\hat{Q})$, is given by $\frac{1}{n} \sum_{i=1}^n \hat{\kappa}_1(X_i) \hat{\pi}(1 | X_i) + \hat{\kappa}_0(X_i) \hat{\pi}(0 | X_i)$. For TMLE based on this formulation, targeting $\hat{\kappa}$ is necessary, unlike in $\psi_2(\hat{Q}^*)$ where $\hat{\gamma}$ was targeted. This also includes targeting $\hat{\mu}$ and $\hat{\pi}$. The goal of targeting $\hat{\kappa}$ is to satisfy $P_n \Phi_M(Q) = o_p(n^{-1/2})$,

where $\Phi_M(Q)(O_i)$ is rewritten in terms of $\kappa_a(X)$ as:

$$\Phi_M(Q)(O_i) = \frac{\mathbb{I}(A_i = a_0)}{\pi(a_0 | X_i)} \left\{ \pi(1 | X_i) \{ \mu(M_i, 1, X_i) - \kappa_1(X_i) \} + \pi(0 | X_i) \{ \mu(M_i, 0, X_i) - \kappa_0(X_i) \} \right\}.$$

Implementing the TMLE $\psi_3(\hat{Q}^*)$ requires iterative updates of $(\hat{\mu}, \hat{\pi}, \hat{\kappa})$, making it more complex than $\psi_2(\hat{Q}^*)$. For practical use, we therefore recommend $\psi_2(\hat{Q}^*)$ due to its simpler implementation.

C.6 TMLE algorithms for estimating the ATE front-door functional

The detailed procedures of constructing a TMLE-based plug-in estimator for $\psi(Q)$ in (1), when M is binary, continuous, or multivariate are shown in Algorithms 1, 2, and 3, respectively.

Algorithm 1 TMLE BASED ON MEDIATOR DENSITY ESTIMATION WITH BINARY M ($\psi_1(\hat{Q}^*)$)

1: **Obtain initial nuisance estimates:** $\hat{\mu}^{(0)}$, $\hat{f}_M^{(0)}$, $\hat{\pi}^{(0)}$, and \hat{p}_X .

Estimate of Q_j at the t^{th} iteration is denoted by $\hat{Q}_j^{(t)}$.

2: **Define loss functions & submodels** indexed by $\varepsilon_A, \varepsilon_M, \varepsilon_Y \in \mathbb{R}$. Given $\hat{Q}^{(t)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$:

- Define the parametric submodels at iteration t as follows:

$$\begin{aligned}\hat{\pi}(\varepsilon_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)})(1 | X) &= \text{expit}\{\text{logit}\{\hat{\pi}^{(t)}(1 | X)\} + \varepsilon_A\{\hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X)\}\}, \\ \hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t)})(1 | A, X) &= \text{expit}\{\text{logit}\{\hat{f}_M^{(t)}(1 | A, X)\} + \varepsilon_M \frac{\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)}{\hat{\pi}^{(t)}(A | X)}\},\end{aligned}$$

$$\hat{\mu}(\varepsilon_Y) = \hat{\mu}^{(t)} + \varepsilon_Y,$$

where $\hat{\eta}^{(t)}(a, X) = \int \hat{\mu}^{(0)}(m, a, X) \hat{f}_M^{(t)}(m | a_0, X) dm$, $\hat{\xi}^{(t)}(m, X) = \sum_{a=0}^1 \hat{\mu}^{(0)}(m, a, X) \hat{\pi}^{(t)}(a | X)$.

- Define the loss functions at iteration t as follows:

$$\begin{aligned}L_A(\tilde{\pi})(O) &= -\log \tilde{\pi}(A | X), \quad L_M(\tilde{f}_M)(O) = -\mathbb{I}(A = a_0) \log \tilde{f}_M(M | A, X), \\ L_Y(\tilde{\mu}; \hat{f}_M^{(t)})(O) &= \{\hat{f}_M^{(t)}(M | a_0, X) / \hat{f}_M^{(t)}(M | A, X)\} \{Y - \tilde{\mu}(M, A, X)\}^2.\end{aligned}$$

3: **Update $\hat{\pi}^{(0)}$ and $\hat{f}_M^{(0)}$ iteratively.** We begin by updating $\hat{\pi}$, though updates can start with either $\hat{\pi}$ or \hat{f}_M . At the t^{th} iteration:

- Given $\hat{Q}^{(t)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$, fit the following logistic regression without an intercept:

$$A \sim \text{offset}(\text{logit } \hat{\pi}^{(t)}(1 | X)) + \hat{H}_A^{(t)}(X), \text{ where } \hat{H}_A^{(t)}(X) := \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X).$$

The coefficient in front of $\hat{H}_A^{(t)}(X)$ is the minimizer $\hat{\varepsilon}_A$. Update $\hat{\pi}^{(t)}$ to $\hat{\pi}^{(t+1)} = \hat{\pi}(\hat{\varepsilon}_A; \hat{\mu}^{(0)}, \hat{f}_M^{(t)})$.

- Given $\hat{Q}^{(\text{temp})} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t+1)}, \hat{p}_X)$, fit the following logistic regression without an intercept:

$$M \sim \text{offset}(\text{logit } \hat{f}_M^{(t)}(1 | a_0, X)) + \hat{H}_M^{(t)}(X), \text{ where } \hat{H}_M^{(t)}(X) := \frac{\hat{\xi}^{(t)}(1, X) - \hat{\xi}^{(t)}(0, X)}{\hat{\pi}^{(t+1)}(a_0 | X)}.$$

Note that $\hat{\xi}^{(t)}$ is computed using $\hat{\mu}^{(0)}$ and $\hat{\pi}^{(t+1)}$.

The coefficient of $\hat{H}_M^{(t)}(X)$ is the minimizer $\hat{\varepsilon}_M$. Update $\hat{f}_M^{(t)}$ to $\hat{f}_M^{(t+1)} = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu}, \hat{\pi}^{(t+1)})$.

- Let $\hat{Q}^{(t+1)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t+1)}, \hat{\pi}^{(t+1)}, \hat{p}_X)$. Iterate over this step while $|\text{P}_n \Phi(\hat{Q}^{(t+1)})| > C_n = o_p(n^{-1/2})$. Assume convergence is achieved at iteration $t = t^*$. Let $\hat{\pi}^* = \hat{\pi}^{(t^*)}$ and $\hat{f}_M^* = \hat{f}_M^{(t^*)}$.

4: **Update $\hat{\mu}^{(0)}$ in one step.**

- Given $\hat{Q}^{(t^*)} = (\hat{\mu}^{(0)}, \hat{f}_M^*, \hat{\pi}^*, \hat{p}_X)$, fit the weighted following regression:

$$Y \sim \text{offset}(\hat{\mu}^{(0)}(M, A, X)) + 1, \text{ with weight } = \hat{f}_M^*(M | a_0, X) / \hat{f}_M^*(M | A, X).$$

The intercept is the minimizer $\hat{\varepsilon}_Y$. Update $\hat{\mu}^{(0)}(M, A, X)$ as $\hat{\mu}^*(M, A, X) = \hat{\mu}^{(0)}(M, A, X) + \hat{\varepsilon}_Y$.

- Let $\hat{Q}^* = (\hat{\mu}^*, \hat{f}_M^*, \hat{\pi}^*, \hat{p}_X)$.

5: **Return** $\psi_1(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}^*(X_i)$ as the TMLE estimator, where

$$\hat{\theta}^*(x) = \sum_{m=0}^1 \hat{\xi}^*(m, x) \hat{f}_M^*(m | a_0, x), \text{ and } \hat{\xi}^*(m, x) = \sum_{a=0}^1 \hat{\mu}^*(m, a, x) \hat{\pi}^*(a | x).$$

Algorithm 2 TMLE BASED ON MEDIATOR DENSITY ESTIMATION WITH CONTINUOUS M ($\psi_1(\hat{Q}^*)$)

1: **Obtain initial nuisance estimates:** $\hat{\mu}^{(0)}$, $\hat{f}_M^{(0)}$, $\hat{\pi}^{(0)}$, and \hat{p}_X .

Estimate of Q_j at the t^{th} iteration is denoted by $\hat{Q}_j^{(t)}$.

2: **Define loss functions & submodels** indexed by $\varepsilon_A, \varepsilon_M, \varepsilon_Y$. Given $\hat{Q}^{(t)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$:

- Define the parametric submodels at iteration t as follows: ($\varepsilon_A, \varepsilon_Y \in \mathbb{R}$, and $-\delta < \varepsilon_M < \delta$)

$$\begin{aligned}\hat{\pi}(\varepsilon_A; \hat{\mu}^{(t)}, \hat{f}_M^{(t)}) (1 | X) &= \text{expit} \left\{ \logit \{ \hat{\pi}^{(t)}(1 | X) \} + \varepsilon_A \{ \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X) \} \right\}, \\ \hat{f}_M(\varepsilon_M; \hat{\mu}^{(t)}, \hat{\pi}^{(t)})(M | a_0, X) &= \hat{f}_M^{(t)}(M | a_0, X) \left\{ 1 + \varepsilon_M \frac{\hat{\xi}^{(t)}(M, X) - \hat{\theta}^{(t)}(X)}{\hat{\pi}^{(t)}(a_0 | X)} \right\},\end{aligned}$$

$$\hat{\mu}(\varepsilon_Y) = \hat{\mu}^{(t)} + \varepsilon_Y,$$

where $\hat{\eta}^{(t)}(a, X) = \int \hat{\mu}^{(t)}(m, a, X) \hat{f}_M^{(t)}(m | a_0, X) dm$, $\hat{\xi}^{(t)}(m, X) = \sum_{a=0}^1 \hat{\mu}^{(t)}(m, a, X) \hat{\pi}^{(t)}(a | X)$.

The parametric submodel for \hat{f}_M can also be chosen to be (45) with $\varepsilon_M \in \mathbb{R}$.

- Define the loss functions at iteration t as follows:

$$\begin{aligned}L_A(\tilde{\pi})(O) &= -\log \tilde{\pi}(A | X), \quad L_M(\tilde{f}_M)(O) = -\mathbb{I}(A = a_0) \log \tilde{f}_M(M | A, X), \\ L_Y(\tilde{\mu}; \hat{f}_M^{(t)})(O) &= \{ \hat{f}_M^{(t)}(M | a_0, X) / \hat{f}_M^{(t)}(M | A, X) \} \{ Y - \tilde{\mu}(M, A, X) \}^2.\end{aligned}$$

3: **Update $\hat{\pi}^{(0)}$ and $\hat{f}_M^{(0)}$ iteratively.** We begin by updating $\hat{\pi}$, though updates can start with either $\hat{\pi}$ or \hat{f}_M . At the t^{th} iteration:

- Given $\hat{Q}^{(t)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t)}, \hat{p}_X)$, fit the following logistic regression without an intercept:

$$A \sim \text{offset}(\logit \hat{\pi}^{(t)}(1 | X)) + \hat{H}_A^{(t)}(X), \text{ where } \hat{H}_A^{(t)}(X) := \hat{\eta}^{(t)}(1, X) - \hat{\eta}^{(t)}(0, X).$$

The coefficient of $\hat{H}_A^{(t)}(X)$ is the minimizer $\hat{\varepsilon}_A$. Update $\hat{\pi}^{(t)}$ to $\hat{\pi}^{(t+1)} = \hat{\pi}(\hat{\varepsilon}_A; \hat{\mu}, \hat{f}_M^{(t)})$.

- Given $\hat{Q}^{(\text{temp})} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t)}, \hat{\pi}^{(t+1)}, \hat{p}_X)$, obtain $\hat{\varepsilon}_M$ by numerically solving this optimization problem:

$$\hat{\varepsilon}_M = \underset{\varepsilon_M \in \mathbb{R}}{\text{argmin}} P_n L_M \left(\hat{f}_M(\varepsilon_M; \hat{\mu}^{(0)}, \hat{\pi}^{(t+1)}) \right).$$

Update $\hat{f}_M^{(t)}$ to $\hat{f}_M^{(t+1)} = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu}, \hat{\pi}^{(t+1)})$.

- Let $\hat{Q}^{(t+1)} = (\hat{\mu}^{(0)}, \hat{f}_M^{(t+1)}, \hat{\pi}^{(t+1)}, \hat{p}_X)$. Iterate over this step while $|P_n \Phi(\hat{Q}^{(t+1)})| > C_n = o_p(n^{-1/2})$.

Assume convergence is achieved at iteration $t = t^*$. Let $\hat{\pi}^* = \hat{\pi}^{(t^*)}$ and $\hat{f}_M^* = \hat{f}_M^{(t^*)}$.

4: **Update $\hat{\mu}^{(0)}$ in one step.**

- Given $\hat{Q}^{(t^*)} = (\hat{\mu}^{(0)}, \hat{f}_M^*, \hat{\pi}^*, \hat{p}_X)$, fit the following weighted regression:

$$Y \sim \text{offset}(\hat{\mu}^{(0)}(M, A, X)) + 1, \text{ with weight } = \hat{f}_M^*(M | a_0, X) / \hat{f}_M^*(M | A, X).$$

The intercept is the minimizer $\hat{\varepsilon}_Y$. Update $\hat{\mu}^{(0)}(M, A, X)$ as $\hat{\mu}^*(M, A, X) = \hat{\mu}^{(0)}(M, A, X) + \hat{\varepsilon}_Y$.

- Let $\hat{Q}^* = (\hat{\mu}^*, \hat{f}_M^*, \hat{\pi}^*, \hat{p}_X)$.

5: **Return** $\psi_1(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}^*(X_i)$ as the TMLE estimator, where

$$\hat{\theta}^*(x) = \int \hat{\xi}^*(m, x) \hat{f}_M^*(m | a_0, x) dm, \text{ and } \hat{\xi}^*(m, x) = \sum_{a=0}^1 \hat{\mu}^*(m, a, x) \hat{\pi}^*(a | x).$$

Algorithm 3 TMLE THAT AVOIDS MEDIATOR DENSITY ESTIMATION ($\psi_2(\hat{Q}^*)$)

- 1: **Obtain initial nuisance estimates:** $\hat{\mu}, \hat{\kappa}_a, \hat{f}_M^r, \hat{\pi}, \hat{\gamma}$, and \hat{p}_X .
 - $\hat{f}_M^r(M, a_1, X)$ can be estimated either via direct estimation of the density ratio, or by applying the Bayes' rule to reparameterize the ratio in terms of $\hat{\pi}(A | X)$ and $\hat{\lambda}(A | M, X)$, as in (16).
 - $\hat{\kappa}_{a_1}(X)$ is obtained via a regression of $\hat{\mu}(M, a, X)$ on X using only rows with $A = a_0$.
- 2: **Define loss functions and parametric fluctuations** indexed by $\varepsilon_A, \varepsilon_\gamma, \varepsilon_Y \in \mathbb{R}$.
 - Define the parametric submodels as follows:

$$\begin{aligned}\hat{\mu}(\varepsilon_Y) &= \hat{\mu} + \varepsilon_Y, \\ \hat{\pi}(\varepsilon_A; \hat{\kappa})(1 | X) &= \text{expit} \left\{ \text{logit} \{ \hat{\pi}(1 | X) \} + \varepsilon_A \{ \hat{\kappa}_1(X) - \hat{\kappa}_0(X) \} \right\}, \\ \hat{\gamma}(\varepsilon_\gamma)(X) &= \hat{\gamma}(X) + \varepsilon_\gamma.\end{aligned}$$
 - Define the loss functions as follows:

$$\begin{aligned}L_Y(\tilde{\mu}; \hat{f}_M^r)(O) &= \hat{f}_M^r(M, A, X) \{Y - \tilde{\mu}(M, A, X)\}^2, \\ L_A(\tilde{\pi})(O) &= -\log \tilde{\pi}(A | X), \\ L_\gamma(\tilde{\gamma}; \hat{\pi}, \hat{\xi})(O) &= \frac{\mathbb{I}(A = a_0)}{\hat{\pi}(a_0 | X)} \left(\hat{\xi}(M, X) - \tilde{\gamma}(X) \right)^2.\end{aligned}$$
- 3: **Update $\hat{\mu}$ and $\hat{\pi}$ in one step** by solving the followings optimization problem:

$$\hat{\varepsilon}_Y = \underset{\varepsilon_Y \in \mathbb{R}}{\text{argmin}} P_n L_Y(\hat{\mu}(\varepsilon_Y); \hat{f}_M^r), \quad \hat{\varepsilon}_A = \underset{\varepsilon_A \in \mathbb{R}}{\text{argmin}} P_n L_A(\hat{\pi}(\varepsilon_A)).$$
 - Fit the following weighted regression and logistic regression without intercept term

$$\begin{aligned}Y &\sim \text{offset}(\hat{\mu}(M, A, X)) + 1, \text{ weight} = \hat{f}_M^r; \\ A &\sim \text{offset}(\text{logit } \hat{\pi}(1 | X)) + \hat{H}_A(X), \quad \text{where } \hat{H}_A(X) = \hat{\kappa}_1(X) - \hat{\kappa}_0(X).\end{aligned}$$
 - $\hat{\varepsilon}_Y$ and $\hat{\varepsilon}_A$ equal the coefficients of the intercept and in front of $\hat{H}_A(X)$, respectively.
 - Update $\hat{\mu}$ and $\hat{\pi}$ as follows

$$\hat{\mu}^* = \hat{\mu}(\hat{\varepsilon}_Y; \hat{f}_M^r), \quad \hat{\pi}^* = \pi(\hat{\varepsilon}_A; \hat{\mu}).$$
 - Define $\hat{\xi}^*(m, x) = \sum_{a=0}^1 \hat{\mu}^*(m, a, x) \hat{\pi}^*(a | x)$. Estimate $\hat{\gamma}(X)$ by fitting the following linear regression using only data points with $A = a_0$:

$$\hat{\xi}^*(m, x) \sim X.$$
- 4: **Update $\hat{\gamma}$ in one step** by solving the followings optimization problem:

$$\hat{\varepsilon}_\gamma = \underset{\varepsilon_\gamma \in \mathbb{R}}{\text{argmin}} P_n L_\gamma \left(\hat{\gamma}(\varepsilon_\gamma); \hat{\pi}^*, \hat{\xi}^* \right).$$
 - Fit the following weighted linear regression

$$\hat{\xi}^* \sim \text{offset}(\hat{\gamma}) + 1, \quad \text{with weight} = \frac{\mathbb{I}(A = a_0)}{\hat{\pi}^*(a_0 | X)}.$$
 - The coefficient of the intercept is $\hat{\varepsilon}_\gamma$, which minimize the empirical risk.
 - Update $\hat{\gamma}(X)$ as $\hat{\gamma}^* = \hat{\gamma}(\hat{\varepsilon}_\gamma)$.
- 5: **Return** $\psi_2(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}^*(X_i)$ as the TMLE estimator.

D Details on estimators for the ATT front-door functional

We first would like to highlight that the ATT estimand in (2) can be redefined as: $\beta(Q) = \psi(Q)/p(A = a_1) - \mathbb{E}(Y | A = a_0)p(A = a_0)$, where $\psi(Q)$ is defined in (1). This reparameterization enables the use of any ATE estimator from Section 3, together with empirical estimates of $\mathbb{E}(Y | A = a_0)$ and $p(A)$, to construct one-step or TMLE estimators for the ATT. While straightforward, this approach introduces unnecessary complexity by estimating nuisance components tailored to the ATE—such as $\hat{\xi}(M, X)$ and $\hat{\theta}(X)$ in $\psi_1(\hat{Q})$, or pseudo-outcome regressions like $\hat{\gamma}(X)$ in $\psi_{2a}(\hat{Q})$ and $\psi_{2b}(\hat{Q})$ —that are irrelevant for ATT estimation. This increases the risk of model misspecification and computational burden. In contrast, the EIF-based estimators proposed in Section 4 target the ATT directly and avoid such extraneous steps.

D.1 Plug-in and one-step ATT estimators under standard factorization

Let $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$ denote the collection of nuisance estimates. When M is discrete, \hat{f}_M can be obtained via regression-based methods; for univariate continuous or mixed-type multivariate mediators, it may be estimated using parametric models, kernel-based approaches, or flexible density estimation techniques [Hayfield and Racine, 2008, Benkeser and Van Der Laan, 2016]. The quantities \hat{p}_A and \hat{p}_{AX} denote the empirical estimates of the marginal and joint distributions of A and (A, X) , respectively. A plug-in estimator of $\beta(Q)$, denoted by $\beta_1(\hat{Q})$, is given by:

$$\beta_1(\hat{Q}) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \int \hat{\mu}(m, a_1, X_i) \hat{f}_M(m | a_0, X_i) dm \right\},$$

where the integral over m simplifies to a summation $\sum_m \hat{\mu}(m, a_1, X_i) \hat{f}_M(m | a_0, X_i)$ when M is discrete, and requires numerical evaluation when M is continuous or of mixed variable types.

The corresponding one-step corrected plug-in estimator, denoted by $\beta_1^+(\hat{Q})$, is given by:

$$\begin{aligned} \beta_1^+(\hat{Q}) &= \beta_1(\hat{Q}) + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M(M_i, a_0, X_i)}{\hat{f}_M(M_i, A_i, X_i)} \{Y_i - \hat{\mu}(M_i, a_1, X_i)\} \right. \\ &\quad + \frac{\mathbb{I}(A_i = a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}(a_1 | X_i)}{\hat{\pi}(a_0 | X_i)} \left\{ \hat{\mu}(M_i, a_1, X_i) - \int \hat{\mu}(m, a_1, X_i) \hat{f}_M(m | a_0, X_i) dm \right\} \\ &\quad \left. + \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \left\{ \int \hat{\mu}(m, a_1, X_i) \hat{f}_M(m | a_0, X_i) dm - \beta_1(\hat{Q}) \right\} \right\}. \end{aligned}$$

To construct the corresponding TMLE, $\beta_1(\hat{Q}^*)$, we follow the same general approach as for the ATE, with details varying by whether M is binary or continuous. The procedures are summarized in Algorithms 4 and 5 in Appendix D.2.

D.2 TMLE algorithms for estimating the ATT front-door functional

The detailed procedures of constructing a TMLE-based plug-in estimator for $\beta(Q)$ in (2), when M is binary, continuous, or multivariate are shown in Algorithms 4, 5, and 6, respectively.

Algorithm 4 TMLE BASED ON MEDIATOR DENSITY ESTIMATION WITH BINARY M ($\beta_1(\hat{Q}^*)$)

- 1: **Obtain initial nuisance estimates:** $\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A$ and \hat{p}_{AX} .
- 2: **Define loss functions & submodels** indexed by $\varepsilon_M, \varepsilon_Y$. Given $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$:

- Define the parametric submodels as follows: ($\varepsilon_M, \varepsilon_Y \in \mathbb{R}$)

$$\hat{f}_M(\varepsilon_M; \hat{\mu})(1 | a_0, X) = \text{expit} \left\{ \text{logit} \left\{ \hat{f}_M(1 | a_0, X) \right\} + \varepsilon_M \frac{\hat{\pi}(a_1 | X)}{\hat{\pi}(a_0 | X)} \frac{\hat{\mu}(1, a_1, X) - \hat{\mu}(0, a_1, X)}{\hat{p}_A(a_1)} \right\},$$

$$\hat{\mu}(\varepsilon_Y)(M, a_1, X) = \hat{\mu}(M, a_1, X) + \varepsilon_Y.$$

- Define the loss functions as follows:

$$L_Y(\tilde{\mu}; \hat{f}_M)(O) = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M(M | a_0, X)}{\hat{f}_M(M | a_1, X)} \{Y - \tilde{\mu}(M, a_1, X)\}^2,$$

$$L_M(\tilde{f}_M)(O) = -\mathbb{I}(A = a_0) \log \tilde{f}_M(M | a_0, X).$$

3: Update \hat{f}_M .

Given $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$, fit the following logistic regression without an intercept:

$$M \sim \text{offset}(\text{logit } \hat{f}_M(1 | a_0, X)) + \hat{H}_M(X), \text{ where } \hat{H}_M(X) := \frac{\hat{\pi}(a_1 | X)}{\hat{\pi}(a_0 | X)} \frac{\hat{\mu}(1, a_1, X) - \hat{\mu}(0, a_1, X)}{\hat{p}_A(a_1)}.$$

- The coefficient in front of $\hat{H}_M(X)$ is the minimizer $\hat{\varepsilon}_M := \text{argmin}_{\varepsilon_M \in \mathbb{R}} P_n L_M(\hat{f}_M(\varepsilon_M; \hat{\mu}))$.
- Update $\hat{f}_M(M | a_0, X)$ to $\hat{f}_M^*(M | a_0, X) = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu})$.

4: Update $\hat{\mu}$.

Given $\hat{Q} = (\hat{\mu}, \hat{f}_M^*, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$, fit the following weighted regression:

$$Y \sim \text{offset}(\hat{\mu}(M, a_1, X)) + 1, \text{ with weight } = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M^*(M | a_0, X)}{\hat{f}_M(M | a_1, X)}.$$

The intercept is the minimizer to $\hat{\varepsilon}_Y := \text{argmin}_{\varepsilon_Y \in \mathbb{R}} P_n L_Y(\hat{\mu}(\varepsilon_Y); \hat{f}_M^*)$.

Update $\hat{\mu}(M, a_1, X)$ as $\hat{\mu}^*(M, a_1, X) = \hat{\mu}(M, a_1, X) + \hat{\varepsilon}_Y$.

5: Return the TMLE estimator $\beta_1(\hat{Q}^*)$ as

$$\beta_1(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \sum_{m \in \{0,1\}} \hat{\mu}^*(m, a_1, X_i) \hat{f}_M^*(m | a_0, X_i).$$

Algorithm 5 TMLE BASED ON MEDIATOR DENSITY ESTIMATION WITH CONTINUOUS M ($\beta_1(\hat{Q}^*)$)

- 1: **Obtain initial nuisance estimates:** $\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A$ and \hat{p}_{AX} .
- 2: **Define loss functions & submodels** indexed by $\varepsilon_M, \varepsilon_Y$. Given $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$:

 - Define the parametric submodels as follows: ($\varepsilon_Y \in \mathbb{R}, \varepsilon_M \in -\delta < \varepsilon_M < \delta$)

$$\hat{f}_M(\varepsilon_M; \hat{\mu})(1 | a_0, X) = \hat{f}_M(1 | a_0, X) \left[1 + \varepsilon_M \left\{ \frac{\hat{\pi}(a_1 | X)}{\hat{\pi}(a_0 | X)} \frac{\hat{\mu}(M, a_1, X) - \hat{k}_{a_1}(X)}{\hat{p}_A(a_1)} \right\} \right],$$

$$\hat{\mu}(\varepsilon_Y)(M, a_1, X) = \hat{\mu}(M, a_1, X) + \varepsilon_Y.$$

The parametric submodel for \hat{f}_M can also be chosen to be (45) with $\varepsilon_M \in \mathbb{R}$.

- Define the loss functions as follows:

$$L_Y(\tilde{\mu}; \hat{f}_M)(O) = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M(M | a_0, X)}{\hat{f}_M(M | a_1, X)} \{Y - \tilde{\mu}(M, a_1, X)\}^2,$$

$$L_M(\tilde{f}_M)(O) = -\mathbb{I}(A = a_0) \log \tilde{f}_M(M | a_0, X).$$

- 3: **Update $\hat{f}_M(M | A, X)$ in one step.**

Given $\hat{Q} = (\hat{\mu}, \hat{f}_M, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$, obtain $\hat{\varepsilon}_M$ by numerically solving this optimization problem:

$$\hat{\varepsilon}_M = \operatorname{argmin}_{\varepsilon_M \in \mathbb{R}} P_n L_M(\hat{f}_M(\varepsilon_M; \hat{\mu})).$$

- Update $\hat{f}_M(M | a_0, X)$ to $\hat{f}_M^*(M | a_0, X) = \hat{f}_M(\hat{\varepsilon}_M; \hat{\mu})$.

- 4: **Update $\hat{\mu}(M, A, X)$ in one step.**

- Given $\hat{Q} = (\hat{\mu}, \hat{f}_M^*, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$, fit the following weighted regression:

$$Y \sim \text{offset}(\hat{\mu}(M, a_1, X)) + 1, \text{ with weight } = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M^*(M | a_0, X)}{\hat{f}_M(M | a_1, X)}.$$

The intercept is the minimizer to $\hat{\varepsilon}_Y = \operatorname{argmin}_{\varepsilon_Y \in \mathbb{R}} P_n L_Y(\hat{\mu}(\varepsilon_Y); \hat{f}_M^*)$.

Update $\hat{\mu}(M, a_1, X)$ as $\hat{\mu}^*(M, a_1, X) = \hat{\mu}(M, a_1, X) + \hat{\varepsilon}_Y$.

- 5: **Return** the TMLE estimator $\beta_1(\hat{Q}^*)$ as

$$\beta_1(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \int \hat{\mu}^*(m, a_1, X_i) \hat{f}_M^*(m | a_0, X_i) dm.$$

Algorithm 6 TMLE THAT AVOIDS MEDIATOR DENSITY ESTIMATION ($\beta(\hat{Q}^*)$)

1: **Obtain initial nuisance estimates:** $\hat{\mu}, \hat{\pi}, \hat{f}_M^r, \hat{\kappa}_{a_1}, \hat{p}_A$, and \hat{p}_{AX} .

- $f_M^r(M, a_1, X)$ can be estimated either via direct estimation of the density ratio, or by applying the Bayes' rule to reparameterize the ratio in terms of $\hat{\pi}(A | X)$ and $\hat{\lambda}(a_1 | M, X)$, as in (16).
- $\hat{\kappa}_{a_1}(X)$ is obtained via a regression of $\hat{\mu}(M, a_1, X)$ on X using only rows with $A = a_0$.

2: **Define loss functions and parametric fluctuations** indexed by ε_κ and ε_Y .

- Define the parametric submodels as follows: ($\varepsilon_Y, \varepsilon_\kappa \in \mathbb{R}$)

$$\hat{\mu}(\varepsilon_Y) = \hat{\mu} + \varepsilon_Y, \quad \hat{\kappa}_{a_1}(\varepsilon_\kappa)(X) = \hat{\kappa}_{a_1}(X) + \varepsilon_\kappa.$$

- Define the loss functions as follows:

$$L_Y(\tilde{\mu}; \hat{f}_M^r)(O) = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \hat{f}_M^r(M, a_1, X) \{Y - \tilde{\mu}(M, a_1, X)\}^2,$$

$$L_\kappa(\tilde{\kappa}_{a_1}; \hat{\pi}, \hat{\mu})(O) = \frac{\mathbb{I}(A = a_0)}{\hat{p}_A(a_1)} (\hat{\mu}(M, a_1, X) - \tilde{\kappa}_{a_1}(X))^2.$$

3: **Update $\hat{\mu}$ in one step.**

Given $\hat{Q} = (\hat{\mu}, \hat{\pi}, \hat{f}_M^r, \hat{\kappa}_{a_1}, \hat{p}_A, \hat{p}_{AX})$, fit the following weighted regression:

$$Y \sim \text{offset}(\hat{\mu}(M, a_1, X)) + 1, \text{ with weight } = \frac{\mathbb{I}(A = a_1)}{\hat{p}_A(a_1)} \hat{f}_M^r(M, a_1, X).$$

- The intercept is the minimizer to $\hat{\varepsilon}_Y = \operatorname{argmin}_{\varepsilon_Y \in \mathbb{R}} P_n L_Y(\hat{\mu}(\varepsilon_Y); \hat{f}_M^r)$.
- Update $\hat{\mu}(M, a_1, X)$ as $\hat{\mu}^*(M, a_1, X) = \hat{\mu}(M, a_1, X) + \hat{\varepsilon}_Y$.
- Estimate $\hat{\kappa}_{a_1}$ by fitting the following linear regression using only data points with $A = a_0$:

$$\hat{\mu}^*(M, a_1, X) \sim X.$$

4: **Update $\hat{\kappa}_{a_1}$ in one step.**

Given $\hat{Q} = (\hat{\mu}^*, \hat{\pi}, \hat{f}_M^r, \hat{\kappa}_{a_1}, \hat{p}_A, \hat{p}_{AX})$, fit the following weighted regression:

$$\hat{\mu}^*(M, a_1, X) \sim \text{offset}(\hat{\kappa}_{a_1}(X)) + 1, \text{ with weight } = \frac{\mathbb{I}(A = a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}(a_1 | X)}{\hat{\pi}(a_0 | X)}.$$

- The intercept is the minimizer to $\hat{\varepsilon}_\kappa = \operatorname{argmin}_{\varepsilon_\kappa \in \mathbb{R}} P_n L_\kappa(\hat{\kappa}_{a_1}(\varepsilon_\kappa); \hat{\pi}, \hat{\mu}^*)$.
- Update $\hat{\kappa}_{a_1}(X)$ as $\hat{\kappa}^*(X) = \hat{\kappa}(X) + \hat{\varepsilon}_\kappa$.

5: **Return** the TMLE estimator $\beta(\hat{Q}^*)$ as

$$\beta(\hat{Q}^*) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(A_i = a_1)}{\hat{p}_A(a_1)} \hat{\kappa}_{a_1}^*(X_i).$$

E Details on inference and asymptotic properties

We assume the following convergence rates for our nuisance estimates:

$$\begin{aligned}
||\hat{\pi}^* - \pi|| &= o_P(n^{-\frac{1}{k}}), & ||\hat{f}_M^* - f_M|| &= o_P(n^{-\frac{1}{b}}), \\
||\hat{\mu}^* - \mu|| &= o_P(n^{-\frac{1}{q}}), & ||\hat{\gamma}^* - \gamma|| &= o_p(n^{-\frac{1}{j}}), \\
||\hat{\kappa}_a - \kappa_a|| &= o_p(n^{-\frac{1}{l}}), & ||\hat{f}_M^r - f_M^r|| &= o_p(n^{-\frac{1}{c}}), \\
||\hat{\lambda} - \lambda|| &= o_P(n^{-\frac{1}{d}}).
\end{aligned} \tag{50}$$

E.1 ATE front-door functional estimators

E.1.1 The remainder term, asymptotic linearity, and robustness for $\psi_1(\hat{Q}^*)$

R₂(\hat{Q}^* , Q) derivation

Given the von Mises expansion, we can write:

$$\begin{aligned}
R_2(\hat{Q}^*, Q) &= \psi(\hat{Q}^*) - \psi(Q) + \int \Phi(\hat{Q}^*) dP(o) \\
&= \iiint \{ \mu(m, a, x) - \hat{\mu}^*(m, a, x) \} \left\{ \frac{\hat{f}_M^*(m | a_0, x)}{\hat{f}_M^*(m | a, x)} f_M(m | a, x) \right\} \pi(a | x) p(x) dx da dm \\
&\quad + \iiint \hat{\mu}^*(m, a, x) \{ f_M(m | a_0, x) - \hat{f}_M^*(m | a_0, x) \} \left\{ \frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} \hat{\pi}^*(a | x) \right\} p(x) dx da dm \\
&\quad + \iiint \left\{ \hat{\mu}^*(m, a, x) \hat{f}_M^*(m | a_0, x) - \mu(m, a, x) f_M(m | a_0, x) \right\} \pi(a | x) p(x) dx da dm.
\end{aligned}$$

To introduce a clear formulation, we introduce a term that is equal to zero into the above expression:

$$\begin{aligned}
0 &= \iiint \frac{f_M(m | a_0, x)}{f_M(m | a, x)} [\mu(m, a, x) - \hat{\mu}(m, a, x)] f_M(m | a, x) \pi(a | x) p(x) dx da dm \\
&\quad + \iiint [\hat{\mu}^*(m, a, x) f_M(m | a_0, x) - \mu(m, a, x) f_M(m | a_0, x)] \pi(a | x) p(x) dx da dm.
\end{aligned}$$

For a more clear derivation of the convergence behavior, we can further decompose $R_2(\hat{Q}^*, Q)$ as:

$$\begin{aligned}
& R_2(\hat{Q}^*, Q) \\
&= \int \left[\frac{\hat{f}_M^*(m | a_0, x)}{\hat{f}_M^*(m | a, x) f_M(m | a, x)} (f_M(m | a, x) - \hat{f}_M^*(m | a, x)) (\mu(m, a, x) - \hat{\mu}^*(m, a, x)) \right. \\
&\quad + \frac{1}{f_M(m | a, x)} (\hat{f}_M^*(m | a_0, x) - f_M(m | a_0, x)) (\mu(m, a, x) - \hat{\mu}^*(m, a, x)) \\
&\quad + \frac{\hat{\mu}^*(m, a, x)}{f_M(m | a, x)} \frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x) \pi(a | x)} (\hat{\pi}^*(a | x) - \pi(a | x)) (f_M(m | a_0, x) - \hat{f}_M^*(m | a_0, x)) \\
&\quad \left. + \frac{\hat{\mu}^*(m, a, x)}{f_M(m | a, x) \hat{\pi}^*(a_0 | x)} (\pi(a_0 | x) - \hat{\pi}^*(a_0 | x)) [f_M(m | a_0, x) - \hat{f}_M^*(m | a_0, x)] \right] dP(x, a, m). \tag{51}
\end{aligned}$$

Regularity discussions

In the following, we discuss two sets of regularity conditions.

[First set of regularity conditions.] Let \mathcal{X} and \mathcal{M} denote the domain of X and M . Assume

$$\begin{aligned}
& \sup_{x \in \mathcal{X}, a \in \{0,1\}, m \in \mathcal{M}} \hat{f}_M^*(m | a, x) / \hat{f}_M^*(m | 1-a, x) < +\infty, \quad \inf_{x \in \mathcal{X}, a \in \{0,1\}} \hat{\pi}^*(a | x) > 0, \\
& \inf_{x \in \mathcal{X}, a \in \{0,1\}, m \in \mathcal{M}} f_M(m | a, x) > 0, \quad \sup_{x \in \mathcal{X}, a \in \{0,1\}} \pi(a | x) / \pi(1-a | x) < +\infty. \tag{52}
\end{aligned}$$

Under the boundedness conditions of (52), we apply the Cauchy–Schwarz inequality to each term in (51), leading to the following inequality:

$$R_2(\hat{Q}^*, Q) \leq C \left[||\hat{f}_M^* - f_M|| \times ||\hat{\mu}^* - \mu|| + ||\hat{f}_M^* - f_M|| \times ||\hat{\pi}^* - \pi|| \right],$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we obtain

$$R_2(\hat{Q}^*, Q) \leq o_P \left(n^{\max\left\{-\left(\frac{1}{b} + \frac{1}{q}\right), -\left(\frac{1}{b} + \frac{1}{k}\right)\right\}} \right). \tag{53}$$

[Second set of regularity conditions.] Let $\|f\|_4 = (\mathbb{P} f^4)^{1/4}$ denote the $L^4(\mathbb{P})$ norm of the function

f . Assume there exists finite constant $C > 0$ such that

$$\left\| \frac{\hat{f}_M^*(\cdot | a_0, \cdot)}{\hat{f}_M^* f_M} \right\|_4 \leq C, \quad \left\| \frac{1}{f_M} \right\|_4 \leq C, \quad \left\| \frac{1}{f_M} \frac{\pi(a_0 | \cdot)}{\hat{\pi}^*(a_0 | \cdot) \pi} \right\|_4 \leq C, \quad \left\| \frac{1}{f_M} \frac{1}{\hat{\pi}^*(a_0 | \cdot)} \right\|_4 \leq C. \quad (54)$$

Given that the boundedness conditions in (54) hold, we apply the Cauchy–Schwarz inequality to each term in (51), resulting in the following inequality:

$$R_2(\hat{Q}^*, Q) \leq C \left[\| \hat{f}_M^* - f_M \|_4 \times \| \hat{\mu}^* - \mu \| + \| \hat{f}_M^* - f_M \| \times \| \hat{\pi}^* - \pi \|_4 \right].$$

We can arrive at the same result as in (53) by modifying the convergence behaviors of \hat{f}_M^* and $\hat{\pi}^*$ in (50) to reflect a stronger $L^4(P)$ -consistency. This can be expressed as follows:

$$\| \hat{\pi}^* - \pi \|_4 = o_P(n^{-\frac{1}{k}}), \quad \| \hat{f}_M^* - f_M \|_4 = o_P(n^{-\frac{1}{b}}). \quad (55)$$

E.1.2 The remainder term, asymptotic linearity, and robustness for $\psi_{2a}(\hat{Q}^*)$

$R_2(\hat{Q}^*, Q)$ derivation

Given the von Mises expansion, we can write:

$$\begin{aligned} R_2(\hat{Q}^*, Q) &= \psi(\hat{Q}^*) - \psi(Q) + \int \Phi(\hat{Q}^*) dP(o) \\ &= \iiint \hat{f}_M^r(m, a, x) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] f_M(m | a, x) \pi(a | x) p(x) dx da dm \\ &\quad + \iint \frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \\ &\quad + \int [\hat{\kappa}_1(x) - \hat{\kappa}_0(x)] (\pi(1 | x) - \hat{\pi}^*(1 | x)) p(x) dx \\ &\quad + \int \hat{\gamma}^*(x) p(x) dx - \int \mathbb{E}(\xi(m, x) | a_0, x) p(x) dx \\ &= \iiint (\hat{f}_M^r(m, a, x) - f_M^r(m, a, x)) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] \\ &\quad \times f_M(m | a, x) \pi(a | x) p(x) dx da dm \\ &\quad + \iiint f_M^r(m, a, x) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] f_M(m | a, x) \pi(a | x) p(x) dx da dm \\ &\quad + \iint \left(\frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} - 1 \right) (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \\ &\quad + \iint (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \end{aligned}$$

$$\begin{aligned}
& + \int \left[(\hat{\kappa}_1(x) - \hat{\kappa}_0(x)) - (\kappa_1(x) - \kappa_0(x)) \right] (\pi(1 | x) - \hat{\pi}^*(1 | x)) p(x) dx \\
& + \int (\kappa_1(x) - \kappa_0(x)) (\pi(1 | x) - \hat{\pi}^*(1 | x)) p(x) dx \\
& + \int \hat{\gamma}^*(x) p(x) dx - \int \mathbb{E}(\xi(m, x) | a_0, x) p(x) dx \\
& = \int \left(\hat{f}_M^r(m, a, x) - f_M^r(m, a, x) \right) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] dP(m, a, x) \\
& + \int \left(\frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} - 1 \right) (\gamma(x)) - \hat{\gamma}^*(x) dP(x) \\
& + \int \left[(\hat{\kappa}_1(x) - \hat{\kappa}_0(x)) - (\kappa_1(x) - \kappa_0(x)) \right] (\pi(1 | x) - \hat{\pi}^*(1 | x)) dP(x).
\end{aligned}$$

Regularity discussions

In the following, we discuss two regularity conditions.

[First regularity condition.] Let \mathcal{X} denote the domain of X . Assume

$$\inf_{x \in \mathcal{X}, a \in \{0, 1\}} \hat{\pi}^*(a | x) > 0. \quad (56)$$

If the condition in (56) holds, then by applying Cauchy–Schwarz inequality, we arrive at the following inequality:

$$\begin{aligned}
R_2(\hat{Q}^*, Q) & \leq \|\hat{f}_M^r - f_M^r\| \times \|\hat{\mu}^* - \mu\| + C \|\hat{\pi}^* - \pi\| \times \|\hat{\gamma}^* - \gamma\| \\
& + \|\hat{\kappa}_1 - \kappa_1\| \times \|\hat{\pi}^* - \pi\| + \|\hat{\kappa}_0 - \kappa_0\| \times \|\hat{\pi}^* - \pi\|,
\end{aligned}$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we have

$$R_2(\hat{Q}^*, Q) \leq o_P \left[n^{\max\left\{-\left(\frac{1}{c} + \frac{1}{q}\right), -\left(\frac{1}{k} + \frac{1}{j}\right), -\left(\frac{1}{k} + \frac{1}{l}\right)\right\}} \right]. \quad (57)$$

[Second set of regularity conditions.] Let $\|f\|_4 = (Pf^4)^{1/4}$ denote the $L^4(P)$ norm of the function f . Assume there exists a finite positive constant C such that

$$\left\| \frac{1}{\hat{\pi}^*} \right\|_4 \leq C. \quad (58)$$

Given the boundedness conditions in (58) hold, we apply the Cauchy-Schwarz inequality to each term in (60), resulting in the following inequality:

$$\begin{aligned} R_2(\hat{Q}^*, Q) &\leq \|\hat{f}_M^r - f_M^r\| \times \|\hat{\mu}^* - \mu\| + C \|\hat{\pi}^* - \pi\|_4 \times \|\hat{\gamma}^* - \gamma\| \\ &\quad + \|\hat{\kappa}_1 - \kappa_1\| \times \|\hat{\pi}^* - \pi\| + \|\hat{\kappa}_0 - \kappa_0\| \times \|\hat{\pi}^* - \pi\|. \end{aligned}$$

We can arrive at the same result as in (57) by modifying the convergence behaviors of $\hat{\pi}^*(A | X)$ in (50) to reflect a stronger $L^4(P)$ -consistency. This can be expressed as follows:

$$\|\hat{\pi}^* - \pi\|_4 = o_P(n^{-\frac{1}{k}}).$$

Remark E.1. It is important to note that the nuisance estimates $\hat{\gamma}^*$ and $\hat{\kappa}_a$ depend on the estimates of $\hat{\xi}^*$ and $\hat{\mu}^*$, respectively. Moreover, $\hat{\xi}^*$ itself relies on the estimates of $\hat{\mu}^*$ and $\hat{\pi}^*$. However, the $L^2(P)$ convergence conditions $\|\hat{\gamma}^* - \gamma\| = o_P(n^{-\frac{1}{j}})$ and $\|\hat{\kappa}_a - \kappa_a\| = o_P(n^{-\frac{1}{l}})$, from display 50, indicate the convergence of the sequential regressions for any choice of $\tilde{\pi} \in \mathcal{M}_\pi$ and $\tilde{\mu} \in \mathcal{M}_\mu$, irrespective of the correctness of these nuisance estimates. To make this dependence more explicit, the respective convergence rates can be restated as follows:

$$\|\hat{\gamma}^*(.; \hat{\mu}^*, \hat{\pi}^*) - \gamma(.; \hat{\mu}^*, \hat{\pi}^*)\| = o_P(n^{-\frac{1}{j}}), \quad \|\hat{\kappa}_a(.; \hat{\mu}^*) - \kappa_a(.; \hat{\mu}^*)\| = o_P(n^{-\frac{1}{l}}). \quad (59)$$

E.1.3 The remainder term, asymptotic linearity, and robustness for $\psi_{2b}(\hat{Q}^*)$

$R_2(\hat{Q}^*, Q)$ derivation

Given the von Mises expansion, we can write:

$$\begin{aligned} R_2(\hat{Q}^*, Q) &= \psi(\hat{Q}^*) - \psi(Q) + \int \Phi(\hat{Q}^*) dP(o) \\ &= \iiint \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)} \frac{\hat{\pi}^*(a | x)}{\hat{\pi}^*(a_0 | x)} [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] f_M(m | a, x) \pi(a | x) p(x) dx da dm \\ &\quad + \iint \frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \\ &\quad + \int [\hat{\kappa}_1(x) - \hat{\kappa}_0(x)] (\pi(1 | x) - \hat{\pi}^*(1 | x)) p(x) dx \end{aligned}$$

$$\begin{aligned}
& + \int \hat{\gamma}^*(x) p(x) dx - \int \mathbb{E}(\xi(m, x) | a_0, x) p(x) dx \\
& = \iiint \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)} \left(\frac{\hat{\pi}(a | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a | x)}{\pi(a_0 | x)} \right) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] \\
& \quad \times f_M(m | a, x) \pi(a | x) p(x) dx da dm \\
& + \iiint \frac{\pi(a | x)}{\pi(a_0 | x)} \left(\frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)} - \frac{\lambda(a_0 | m, x)}{\lambda(a | m, x)} \right) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] \\
& \quad \times f_M(m | a, x) \pi(a | x) p(x) dx da dm \\
& + \iiint f_M^r(m, a, x) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] f_M(m | a, x) \pi(a | x) p(x) dx da \\
& + \iint \left(\frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} - 1 \right) (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \\
& + \iint (\hat{\xi}^*(m, x) - \hat{\gamma}^*(x)) f_M(m | a_0, x) p(x) dx dm \\
& + \int \left[(\hat{\kappa}_1(x) - \hat{\kappa}_0(x)) - (\kappa_1(x) - \kappa_0(x)) \right] (\pi(1 | x) - \hat{\pi}(1 | x)) p(x) dx \\
& + \int (\kappa_1(x) - \kappa_0(x)) (\pi(1 | x) - \hat{\pi}^*(1 | x)) p(x) dx \\
& = \int \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)} \left(\frac{\hat{\pi}^*(a | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a | x)}{\pi(a_0 | x)} \right) [\mu(m, a, x) - \hat{\mu}(m, a, x)] dP(m, a, x) \\
& + \int \frac{\pi(a | x)}{\pi(a_0 | x)} \left(\frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)} - \frac{\lambda(a_0 | m, x)}{\lambda(a | m, x)} \right) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] dP(m, a, x) \\
& + \int \left(\frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} - 1 \right) (\gamma(x)) - \hat{\gamma}^*(x) dP(x) \\
& + \int \left[(\hat{\kappa}_1(x) - \hat{\kappa}_0(x)) - (\kappa_1(x) - \kappa_0(x)) \right] (\pi(1 | x) - \hat{\pi}^*(1 | x)) dP(x).
\end{aligned}$$

For a clearer derivation of the convergence behavior, we can further decompose $R_2(\hat{Q}^*, Q)$ as:

$$\begin{aligned}
& R_2(\hat{Q}^*, Q) \\
&= \int \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)\hat{\pi}^*(a_0 | x)} (\hat{\pi}^*(a | x) - \pi(a | x)) [\mu(m, a, x) - \hat{\mu}(m, a, x)] dP(m, a, x) \\
&\quad + \int \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a | m, x)\hat{\pi}^*(a_0 | x)} \frac{\pi(a | x)}{\hat{\pi}^*(a_0 | x)\pi(a_0 | x)} \\
&\quad \quad \times (\pi(a_0 | x) - \hat{\pi}^*(a_0 | x)) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] dP(m, a, x) \\
&\quad + \int \frac{\pi(a | x)}{\pi(a_0 | x)\hat{\lambda}(a_0 | m, x)} (\hat{\lambda}(a | m, x) - \lambda(a | m, x)) [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] dP(m, a, x) \\
&\quad + \int \frac{\pi(a | x)}{\pi(a_0 | x)} \frac{\lambda(a | m, x)}{\hat{\lambda}(a_0 | m, x)\lambda(a_0 | m, x)} (\lambda(a_0 | m, x) - \hat{\lambda}(a_0 | m, x)) \\
&\quad \quad \times [\mu(m, a, x) - \hat{\mu}^*(m, a, x)] dP(m, a, x) \\
&\quad + \int \left(\frac{\pi(a_0 | x)}{\hat{\pi}^*(a_0 | x)} - 1 \right) (\gamma(x)) - \hat{\gamma}^*(x) dP(x) \\
&\quad + \int [(\hat{\kappa}_1(x) - \hat{\kappa}_0(x)) - (\kappa_1(x) - \kappa_0(x))] (\pi(1 | x) - \hat{\pi}^*(1 | x)) dP(x). \tag{60}
\end{aligned}$$

Regularity discussions

In the following, we discuss two sets of regularity conditions.

[First set of regularity conditions.] Let \mathcal{X} and \mathcal{M} denote the domain of X and M . Assume

$$\begin{aligned}
& \inf_{a \in \{0,1\}, x \in \mathcal{X}} \hat{\pi}^*(a | x) > 0, \quad \sup_{x \in \mathcal{X}, a \in \{0,1\}, m \in \mathcal{M}} \hat{\lambda}(a | x, m) / \hat{\lambda}(1-a | x, m) < +\infty, \\
& \sup_{x \in \mathcal{X}, a \in \{0,1\}} \pi(a | x) / \pi(1-a | x) < +\infty, \quad \inf_{x \in \mathcal{X}, a \in \{0,1\}, m \in \mathcal{M}} \hat{\lambda}(a | m, x) > 0. \tag{61}
\end{aligned}$$

Under the boundedness conditions of (61), we apply the Cauchy–Schwarz inequality to each term in (60), leading to the following inequality:

$$\begin{aligned}
R_2(\hat{Q}^*, Q) &\leq C \left[||\hat{\pi}^* - \pi|| \times ||\hat{\mu}^* - \mu|| + ||\hat{\lambda} - \lambda|| \times ||\hat{\mu}^* - \mu|| \right] \\
&\quad + ||\hat{\pi}^* - \pi|| \times ||\hat{\gamma}^* - \gamma|| + ||(\hat{\kappa}_1 - \hat{\kappa}_0) - (\kappa_1 - \kappa_0)|| \times ||\hat{\pi}^* - \pi|| \right],
\end{aligned}$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we obtain

$$R_2(\hat{Q}^*, Q) \leq o_P \left[n^{\max\left\{-\left(\frac{1}{q} + \frac{1}{k}\right), -\left(\frac{1}{d} + \frac{1}{q}\right), -\left(\frac{1}{k} + \frac{1}{j}\right), -\left(\frac{1}{k} + \frac{1}{\ell}\right)\right\}} \right]. \quad (62)$$

[Second set of regularity conditions.] Let $\|f\|_4 = (Pf^4)^{1/4}$ denote the $L^4(P)$ norm of the function f . Assume there exists a finite positive constant C such that

$$\begin{aligned} \left\| \frac{\hat{\lambda}(a_0 | .)}{\hat{\lambda}\hat{\pi}^*(a_0 | .)} \right\|_4 &\leq C, & \left\| \frac{\hat{\lambda}(a_0 | .)}{\hat{\lambda}\hat{\pi}^*(a_0 | .)} \frac{\pi}{\hat{\pi}^*(a_0 | X)} \frac{\pi(a_0 | .)}{\pi(a_0 | .)} \right\|_4 &\leq C, \\ \left\| \frac{\pi}{\pi(a_0 | .)\hat{\lambda}(a_0 | .)} \right\|_4 &\leq C, & \left\| \left(\frac{\pi(a_0 | .)}{\hat{\pi}^*(a_0 | .)} - 1 \right) \right\|_4 &\leq C. \end{aligned} \quad (63)$$

Given that the boundedness conditions in (63) hold, we apply the Cauchy-Schwarz inequality to each term in (60), resulting in the following inequality:

$$\begin{aligned} R_2(\hat{Q}^*, Q) &\leq C \left[\|\hat{\pi}^* - \pi\|_4 \times \|\hat{\mu} - \mu\| + \|\hat{\lambda} - \lambda\|_4 \times \|\hat{\mu}^* - \mu\| \right] \\ &\quad + \|\hat{\pi}^* - \pi\|_4 \times \|\hat{\gamma}^* - \gamma\| + \|(\hat{\kappa}_1 - \hat{\kappa}_0) - (\kappa_1 - \kappa_0)\| \times \|\hat{\pi}^* - \pi\|. \end{aligned}$$

We can arrive at the same result as in (62) by modifying the convergence behaviors of $\hat{\lambda}(A | M, X)$ and $\hat{\pi}^*(1 | X)$ in (50) to reflect a stronger $L^4(P)$ -consistency. This can be expressed as follows:

$$\|\hat{\pi}^* - \pi\|_4 = o_P(n^{-\frac{1}{k}}), \quad \|\hat{\lambda} - \lambda\|_4 = o_P(n^{-\frac{1}{d}}).$$

E.2 ATT front-door functional estimators

E.2.1 The remainder term, asymptotic linearity, and robustness for $\beta_1(\hat{Q}^*)$

$R_2(\hat{Q}^*, Q)$ derivation

Given the von Mises expansion, we can write:

$$R_2(\hat{Q}^*, Q) = \beta_1(\hat{Q}^*) - \beta_1(Q) + \int \Phi_\beta(\hat{Q}^*) dP(o)$$

$$\begin{aligned}
&= \beta_1(\hat{Q}^*) - \beta_1(Q) + \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{\hat{f}_M(m, a_0, x)}{\hat{f}_M(m, a_1, x)} \{y - \hat{\mu}^*(m, a_1, x)\} \right. \\
&\quad + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \frac{\hat{\pi}^*(a_1 | x)}{\hat{\pi}^*(a_0 | x)} \{\hat{\mu}(m, a_1, x) - \hat{\kappa}_{a_1}(x)\} + \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \{\hat{\kappa}_{a_1}(X) - \beta(\hat{Q}^*)\} \Big\} dP(o) \\
&= \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \left(\frac{\hat{f}_M(m, a_0, x)}{\hat{f}_M(m, a_1, x)} - \frac{f_M(m, a_0, x)}{f_M(m, a_1, x)} \right) \{\mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x)\} \right. \\
&\quad + \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{f_M(m, a_0, x)}{f_M(m, a_1, x)} \{\mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x)\} \\
&\quad + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \left(\frac{\hat{\pi}^*(a_1 | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \right) \hat{\mu}^*(m, a_1, x) \{f_M(m, a_0, x) - \hat{f}_M(m, a_0, x)\} \\
&\quad + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \hat{\mu}^*(m, a_1, x) \{f_M(m, a_0, x) - \hat{f}_M(m, a_0, x)\} \\
&\quad \left. + \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \{\hat{\kappa}_{a_1}(X) - \beta(\hat{Q}^*)\} \right\} dP(o) \\
&\quad + \beta_1(\hat{Q}^*) - \beta_1(Q).
\end{aligned}$$

Note that the second, fourth, fifth, and sixth lines sum to a term with $o_p(n^{-\frac{1}{2}})$ rate of convergence:

$$\begin{aligned}
&(2) + (4) + (5 - 6) \\
&= \beta_1(\hat{Q}^*) - \beta_1(Q) \\
&+ \int \left\{ \underbrace{\frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{f_M(m, a_0, x)}{f_M(m, a_1, x)} \mu(m, a_1, x)}_{\textcircled{1}} - \underbrace{\frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{f_M(m, a_0, x)}{f_M(m, a_1, x)} \hat{\mu}^*(m, a_1, x)}_{\textcircled{2}} \right. \\
&\quad + \underbrace{\frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \hat{\mu}^*(m, a_1, x) f_M(m, a_0, x)}_{\textcircled{3}} - \underbrace{\frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \hat{\mu}^*(m, a_1, x) \hat{f}_M(m, a_0, x)}_{\textcircled{4}} \\
&\quad \left. + \underbrace{\frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \hat{\kappa}_{a_1}(X)}_{\textcircled{5}} - \underbrace{\frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \beta(\hat{Q}^*)}_{\textcircled{6}} \right\} dP(o),
\end{aligned}$$

where

$$\begin{aligned}
\int \textcircled{2} + \textcircled{3} dP(o) &= \int \textcircled{4} + \textcircled{5} dP(o) = 0, \\
\int \textcircled{1} dP(o) - \beta(Q) &= \int \mathbb{I}(a = a_1) \left(\frac{1}{\hat{p}_A(a_1)} - \frac{1}{p_A(a_1)} \right) f_M(m, a_0, x) \mu(m, a_1, x) dP(o), \\
\beta(\hat{Q}^*) - \int \textcircled{6} dP(o) &= (1 - \frac{p_A(a_1)}{\hat{p}_A(a_1)}) \beta(\hat{Q}^*).
\end{aligned}$$

Therefore, we have the second-order remainder term to have the final form:

$$\begin{aligned}
R_2(\hat{Q}^*, Q) &= \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \left(\frac{\hat{f}_M(m, a_0, x)}{\hat{f}_M(m, a_1, x)} - \frac{f_M(m, a_0, x)}{f_M(m, a_1, x)} \right) \{\mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x)\} \right. \\
&\quad + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \left(\frac{\hat{\pi}^*(a_1 | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \right) \hat{\mu}^*(m, a_1, x) \{f_M(m, a_0, x) - \hat{f}_M(m, a_0, x)\} \Big\} dP(o) \\
&\quad + \int \mathbb{I}(a = a_1) \left(\frac{1}{\hat{p}_A(a_1)} - \frac{1}{p_A(a_1)} \right) f_M(m, a_0, x) \mu(m, a_1, x) dP(o) \\
&\quad + (1 - \frac{p_A(a_1)}{\hat{p}_A(a_1)}) \beta_1(\hat{Q}^*).
\end{aligned} \tag{64}$$

Regularity discussions

Let \mathcal{X} and \mathcal{M} denote the domain of X and M . Assume

$$\begin{aligned}
\inf_{x \in \mathcal{X}, m \in \mathcal{M}} \hat{f}_M(m, a_1, x) &> 0, & \inf_{x \in \mathcal{X}, m \in \mathcal{M}} f_M(m, a_1, x) &> 0, \\
\inf_{x \in \mathcal{X}} \hat{\pi}(a_0 | x) &> 0, & \inf_{x \in \mathcal{X}} \pi(a_0 | x) &> 0, \\
\sup_{x \in \mathcal{X}, m \in \mathcal{M}} \hat{\mu}^*(m, a_1, x) &< \infty.
\end{aligned} \tag{65}$$

Under the boundedness conditions of (65), we apply the Cauchy–Schwarz inequality to each term in (64), leading to the following inequality:

$$R_2(\hat{Q}^*, Q) \leq C \left[\|\hat{f}_M - f_M\| \times \|\hat{\mu}^* - \mu\| + \|\hat{f}_M - f_M\| \times \|\hat{\pi}^* - \pi\| \right],$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we obtain

$$R_2(\hat{Q}^*, Q) \leq o_P \left[n^{\max\left\{-\left(\frac{1}{b} + \frac{1}{k}\right), -\left(\frac{1}{k} + \frac{1}{q}\right)\right\}} \right]. \tag{66}$$

Asymptotic linearity

Theorem E.2 (Asymptotic linearity of $\beta_1(\hat{Q}^*)$). *In addition to (A1)-(A3) and the boundedness condition (65), we assume that the nuisance estimates $\hat{Q}^* = (\hat{\mu}^*, \hat{f}_M^*, \hat{\pi}, \hat{p}_A, \hat{p}_{AX})$ satisfy:*

(A5.4) $L^2(P)$ convergence of nuisance regressions: Let $\|\hat{\pi}^* - \pi\| = o_P(n^{-\frac{1}{k}})$, $\|\hat{f}_M^* - f_M\| = o_P(n^{-\frac{1}{b}})$, $\|\hat{\mu}^* - \mu\| = o_P(n^{-\frac{1}{q}})$, and assume that both $\frac{1}{b} + \frac{1}{q} \geq \frac{1}{2}$ and $\frac{1}{k} + \frac{1}{b} \geq \frac{1}{2}$.

Under these conditions, $\beta_1(\hat{Q}^*) - \beta_1(Q) = P_n \Phi_\beta(Q) + o_P(n^{-1/2})$ implying that the TMLE $\beta_1(\hat{Q}^*)$ is asymptotically linear and with influence function equal to $\Phi_\beta(Q)$.

Note that \hat{p}_A is estimated nonparametrically as the sample mean. Therefore, it converges to the true mean at a rate of $o_p(n^{-1/2})$, ensuring the last two lines in $R_2(\hat{Q}^*, Q)$ also converge to the truth at a rate of $o_p(n^{-1/2})$.

An immediate corollary of Theorem E.2 is that $\beta_1(\hat{Q}^*)$ shares the same multiple robustness properties as its corresponding ATE estimator, as stated in the Corollary 5.2. To avoid redundancy, we omit a restatement of these properties here.

E.2.2 The remainder term, asymptotic linearity, and robustness for $\beta_a(\hat{Q}^*)$

$R_2(\hat{Q}^*, Q)$ derivation

Given the $R_2(\hat{Q}^*, Q)$ term of $\beta_1(\hat{Q}^*)$, it immediately follows that $\beta_a(\hat{Q}^*)$ has an $R_2(\hat{Q}^*, Q)$ term as follows:

$$\begin{aligned} R_2(\hat{Q}^*, Q) &= \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} (\hat{f}_M^r(m, a, x) - f_M^r(m, a, x)) \{ \mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x) \} \right. \\ &\quad + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \left(\frac{\hat{\pi}^*(a_1 | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \right) \{ \kappa_{a_1}(x; \hat{\mu}^*) - \hat{\kappa}_{a_1}(x; \hat{\mu}^*) \} \Big\} dP(o) \\ &\quad + \int \mathbb{I}(a = a_1) \left(\frac{1}{\hat{p}_A(a_1)} - \frac{1}{p_A(a_1)} \right) f_M(m, a_0, x) \mu(m, a_1, x) dP(o) \\ &\quad + (1 - \frac{p_A(a_1)}{\hat{p}_A(a_1)}) \beta_a(\hat{Q}^*). \end{aligned} \tag{67}$$

Regularity discussions

Let \mathcal{X} and \mathcal{M} denote the domain of X and M . Assume

$$\inf_{x \in \mathcal{X}} \hat{\pi}(a_0 | x) > 0, \quad \inf_{x \in \mathcal{X}} \pi(a_0 | x) > 0. \tag{68}$$

Under the boundedness conditions of (68), we apply the Cauchy–Schwarz inequality to each term in (67), leading to the following inequality:

$$R_2(\hat{Q}^*, Q) \leq C \left[\| \hat{f}_M^r - f_M \| \times \| \hat{\mu}^* - \mu \| + \| \hat{\pi} - \pi \| \times \| \hat{\kappa}_{a_1} - \kappa_{a_1} \| \right],$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we obtain

$$R_2(\hat{Q}^*, Q) \leq o_p \left[n^{\max\{-\left(\frac{1}{c} + \frac{1}{q}\right), -\left(\frac{1}{k} + \frac{1}{l}\right)\}} \right]. \tag{69}$$

Asymptotic linearity

Theorem E.3 (Asymptotic linearity of $\beta_a(\hat{Q}^*)$). *In addition to (A1)-(A3) and the regularity conditions (68), we assume the nuisance estimates $\hat{Q}^* = (\hat{\mu}^*, \hat{\kappa}_{a_1}, \hat{f}_M^r, \hat{\pi}^*, \hat{p}_A, \hat{p}_{AX})$ satisfy:*

(A5.5) *$L^2(P)$ -rates of nuisance estimates: Let $\|\hat{\pi}^* - \pi\| = o_P(n^{-\frac{1}{k}})$, $\|\hat{\mu}^* - \mu\| = o_P(n^{-\frac{1}{q}})$, $\|\hat{\kappa}_{a_1} - \kappa_{a_1}\| = o_P(n^{-\frac{1}{\ell}})$, $\|\hat{f}_M^r - f_M^r\| = o_P(n^{-\frac{1}{c}})$, and assume that $\frac{1}{c} + \frac{1}{q} \geq \frac{1}{2}$, and $\frac{1}{\ell} + \frac{1}{k} \geq \frac{1}{2}$*

Under these conditions, $\beta_a(\hat{Q}^*) - \beta_a(Q) = P_n \Phi_\beta(Q) + o_P(n^{-1/2})$ implying that the TMLE $\beta_a(\hat{Q}^*)$ is asymptotically linear and with influence function equal to $\Phi_\beta(Q)$.

Comparing the asymptotic behavior of $\beta_a(\hat{Q}^*)$ with its ATE counterpart, we find that $\beta_a(\hat{Q}^*)$ demonstrates greater robustness. Specifically, the convergence of $\hat{\gamma}^*$ to its truth at a certain rate—required for $\psi_{2a}(\hat{Q}^*)$ to achieve asymptotic linearity—is no longer necessary for $\beta_a(\hat{Q}^*)$. For the same reason, $\beta_a(\hat{Q}^*)$ achieves consistency under weaker conditions, as illustrated below.

Corollary E.4 (Robustness of $\beta_a(\hat{Q}^*)$). *$\beta_a(\hat{Q}^*)$ is consistent for $\beta(Q)$ if at least one of the following conditions hold:*

- (i) $\|\hat{\pi}^* - \pi\| = o_P(1)$ and $\|\hat{\mu}^* - \mu\| = o_P(1)$,
- (ii) $\|\hat{\pi}^* - \pi\| = o_P(1)$ and $\|\hat{f}_M^r - f_M^r\| = o_P(1)$,
- (iii) $\|\hat{\mu}^* - \mu\| = o_P(1)$, and $\|\hat{\kappa}_1 - \kappa_1\| = o_P(1)$,
- (iv) $\|\hat{\kappa}_{a_1} - \kappa_{a_1}\| = o_P(1)$, and $\|\hat{f}_M^r - f_M^r\| = o_P(1)$.

Corollary E.4 suggests that either the nuisance estimates $\hat{\mu}^*$ and $\hat{\pi}^*$ need to converge to their respective truths (conditions (i)-(iii)) or the estimates introduced to circumvent density estimation, $\hat{\kappa}_{a_1}, \hat{f}_M^r$, should converge to their true values (condition (iv)).

E.2.3 The remainder term, asymptotic linearity, and robustness for $\beta_b(\hat{Q}^*)$

$R_2(\hat{Q}^*, Q)$ derivation

Given the $R_2(\hat{Q}^*, Q)$ term for $\beta_1(\hat{Q}^*)$, it immediately follows that $\beta_b(\hat{Q}^*)$ has an $R_2(\hat{Q}^*, Q)$

term as follows:

$$\begin{aligned}
& R_2(\hat{Q}^*, Q) \\
&= \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \left(\frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a_1 | m, x)} / \frac{\hat{\pi}(a_0 | x)}{\hat{\pi}(a_1 | x)} - \frac{\lambda(a_0 | m, x)}{\lambda(a_1 | m, x)} / \frac{\pi(a_0 | x)}{\pi(a_1 | x)} \right) \{ \mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x) \} \right. \\
&\quad \left. + \frac{\mathbb{I}(a = a_0)}{\hat{p}_A(a_1)} \left(\frac{\hat{\pi}^*(a_1 | x)}{\hat{\pi}^*(a_0 | x)} - \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \right) \{ \kappa_{a_1}(x; \hat{\mu}^*) - \hat{\kappa}_{a_1}(x; \hat{\mu}^*) \} \right\} dP(o) \\
&\quad + \int \mathbb{I}(a = a_1) \left(\frac{1}{\hat{p}_A(a_1)} - \frac{1}{p_A(a_1)} \right) f_M(m, a_0, x) \mu(m, a_1, x) dP(o) \\
&\quad + (1 - \frac{p_A(a_1)}{\hat{p}_A(a_1)}) \beta_a(\hat{Q}^*),
\end{aligned} \tag{70}$$

where the first line can be further simplified as

$$\begin{aligned}
(1) &= \int \left\{ \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{\pi(a_1 | x)}{\pi(a_0 | x)} \left(\frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a_1 | m, x)} - \frac{\lambda(a_0 | m, x)}{\lambda(a_1 | m, x)} \right) \{ \mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x) \} \right. \\
&\quad \left. + \frac{\mathbb{I}(a = a_1)}{\hat{p}_A(a_1)} \frac{\hat{\lambda}(a_0 | m, x)}{\hat{\lambda}(a_1 | m, x)} \left(\frac{\hat{\pi}(a_0 | x)}{\hat{\pi}(a_1 | x)} - \frac{\pi(a_0 | x)}{\pi(a_1 | x)} \right) \{ \mu(m, a_1, x) - \hat{\mu}^*(m, a_1, x) \} \right\} dP(o).
\end{aligned}$$

Regularity discussions

Let \mathcal{X} and \mathcal{M} denote the domain of X and M . Assume

$$\begin{aligned}
\inf_{x \in \mathcal{X}} \hat{\pi}(a_0 | x) &> 0, & \inf_{x \in \mathcal{X}} \pi(a_0 | x) &> 0, \\
\inf_{x \in \mathcal{X}, m \in \mathcal{M}} \hat{\lambda}(a_1 | m, x) &> 0, & \inf_{x \in \mathcal{X}, m \in \mathcal{M}} \lambda(a_1 | m, x) &> 0.
\end{aligned} \tag{71}$$

Under the boundedness conditions of (71), we apply the Cauchy–Schwarz inequality to each term in (70), leading to the following inequality:

$$R_2(\hat{Q}^*, Q) \leq C \left[\|\hat{\pi} - \pi\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\lambda} - \lambda\| \times \|\hat{\mu}^* - \mu\| + \|\hat{\pi} - \pi\| \times \|\hat{\kappa}_{a_1} - \kappa_{a_1}\| \right],$$

where C is a finite positive constant. Given the nuisance convergence rates in (50), we obtain

$$R_2(\hat{Q}^*, Q) \leq o_P \left[n^{\max \left\{ -\left(\frac{1}{k} + \frac{1}{q} \right), -\left(\frac{1}{d} + \frac{1}{q} \right), -\left(\frac{1}{k} + \frac{1}{\ell} \right) \right\}} \right]. \tag{72}$$



Figure 2: Two variations of the front-door graph incorporating an anchor variable Z . At least one of the dashed edges between Z and A must be present to satisfy the *relevance* criterion. This means Z may influence A either directly ($Z \rightarrow A$), through unmeasured confounding ($Z \leftrightarrow A$), or both. Additionally, (a) Z may directly affect the mediator M ($Z \rightarrow M$); or (b) Z may be confounded with M via unmeasured factors ($Z \leftrightarrow M$).

F Details on model evaluation and efficiency gains

To complement the front-door identification and estimation strategies, we expand here on how the presence of an anchor variable Z can support both model evaluation and efficiency gains. This extends our discussion in Section 6 by formalizing the statistical tests and empirical criteria that can be applied when an anchor is available.

F.1 Details on Verma constraint

Bhattacharya and Nabi [2022] introduced the concept of an anchor variable Z to facilitate empirical evaluation of the front-door assumptions. An anchor satisfies the *relevance* criterion, meaning it must be associated with the treatment A , either via a direct effect, through unmeasured confounding, or both. The extended front-door model incorporating such an anchor is shown in Fig. 2, where bidirected arrows indicate the presence of unmeasured confounding between endpoint variables. The anchor variable may also exhibit marginal dependence with the mediator M , either through a direct causal link ($Z \rightarrow M$), as in Fig. 2(a), or via unmeasured confounding ($Z \leftrightarrow M$), as in Fig. 2(b).

In the anchor-included front-door models shown in Fig. 2, Z is marginally associated with Y , even though these variables are not adjacent; that is, $Z \not\perp\!\!\!\perp Y \mid X, A, M$, or conditional on any subset of X, A, M . According to d-separation rules [Pearl, 2009], this dependence arises due to an open path from Z to Y through A and M , which, when blocked (e.g., by conditioning on either A or M), opens up a collider path via the unmeasured confounder between treatment and outcome.

Despite the lack of ordinary conditional independence between Z and Y , their non-adjacency induces a constraint on the observed distribution $P(O)$ where $O = (X, Z, A, M, Y)$, formalized by the nested Markov model for DAGs with hidden variables [Richardson et al., 2023]. For



Figure 3: (a) An example of an anchor-included front-door graph; (b) The conditional graph corresponding to the kernel $q_{AY}(A, Y | X, Z, M)$, where $\{X, Z, M\}$ are fixed by intervention (indicated by square nodes).

concreteness and without loss of generality, we focus on an anchor configuration consistent with the structure in Fig. 3(a).

According to the nested Markov factorization [Richardson et al., 2023], the observed data distribution $P(O)$ for the graph in Fig. 3(a) factorizes as follows:

$$P(X, Z, A, M, Y) = q_X(X) \times q_Z(Z | X) \times q_M(M | X, Z, A) \times q_{AY}(A, Y | X, Z, M),$$

where $q_D(D | \text{pa}_G(D))$ denotes a Markov kernel. Here, D is a *district* which is a set of variables connected by bidirected edges, and the kernel corresponds to a post-intervention distribution in which all variables in $O \setminus D$ are fixed by intervention. Each kernel is identifiable from $P(O)$ via sequential application of the g-formula. In this example, we have $q_X(X) \equiv P(X)$, $q_Z(Z | X) \equiv P(Z | X)$, $q_M(M | X, Z, A) \equiv P(M | X, Z, A)$, and $q_{AY}(A, Y | X, Z, M) \equiv P(A | X, Z) \times P(Y | X, Z, A, M)$.

The kernel $q_{AY}(A, Y | X, Z, M)$ corresponds to the conditional graph in Fig. 3(b), where the variables X , Z , and M are treated as fixed (i.e., all incoming edges into these nodes are removed), as indicated by the square boxes around them. In this conditional graph, Y is d-separated from Z given $\{X, M\}$, implying the independence $Y \perp Z | X, M$. This independence is encoded in the marginal kernel $q_{AY}(Y | X, Z, M)$, which therefore must not depend on Z . Applying the rules of marginalization to $q_{AY}(A, Y | X, Z, M)$ yields the following constraint:

$$q_{AY}(Y | X, Z, M) := \sum_{a'} P(A = a' | X, Z) P(Y | X, Z, A = a', M) \text{ is not a function of } Z. \quad (73)$$

This restriction is an example of a *generalized* independence, also known as a *Verma*, constraint.

Per the results of Tian and Pearl [2002], the post-intervention distributions $P(X, Z, A, Y^m)$ and $P(X, Z, M^a, Y^a)$ are both identifiable, since M and A satisfy the graphical condition of primal

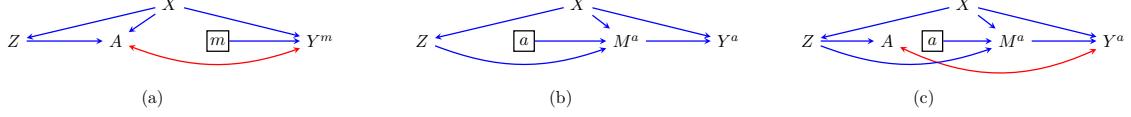


Figure 4: (a) Fixing $M = m$ induces the independence $Z \perp Y^m | X$ in $P(X, Z, A, Y^m)$; (b) Fixing $A = a$ induces the independence $Z \perp Y^a | X, M^a$ in $P(X, Z, M^a, Y^a)$; (c) The graph corresponding to $P(X, Z, A, M^a, Y^a)$.

fixability [Bhattacharya et al., 2022]. A variable $O_i \in O$ is said to be primal fixable if it does not have a path to any of its children that passes only through unmeasured variables. The identified forms of these distributions are as follows:

$$\begin{aligned} P(X, Z, A, Y^m) &:= \frac{P(O)}{q_M(M | X, Z, A)} \Big|_{M=m} = \frac{P(X, Z, A, M=m, Y)}{P(M=m | A, Z, X)} \\ &= P(X, Z, A) \times P(Y | X, Z, A, M=m) , \end{aligned} \quad (74)$$

which is Markov relative to the graph in Fig. 4(a), where $Z \perp Y^m | X$. Similarly,

$$\begin{aligned} P(X, Z, M^a, Y^a) &:= \frac{P(O)}{q_{AY}(A | X, Z, M, Y)} \Big|_{A=a} = \frac{P(X, Z, A=a, M, Y)}{\sum_{a'} P(A=a' | X, Z) P(Y | X, Z, A=a', M)} \\ &= P(X, Z) \times P(M | X, Z, A=a) \times \sum_{a'} P(A=a' | X, Z) \times P(Y | X, Z, A=a', M) , \end{aligned} \quad (75)$$

which is Markov relative to the graph in Fig. 4(b), where $Y^a \perp Z | X, M^a$.

The independencies between counterfactual and factual variables shown in Fig. 4 represent two equivalent forms of the Verma constraint in (73), which underlie our proposed testing procedures. Our weighted risk minimization tests are designed to assess these equivalent independence conditions: the dual test targets the independence in Fig. 4(a), while the primal test targets the one in Fig. 4(b).

Given the identification of $P(X, Z, M^a, Y^a)$ in (75), we can thus write the risk minimizer in (31) in terms of observed data via (32), where $q_{\text{primal}}(A | Y, M, Z, X)$ is simply $1/q_{AY}(A | X, Z, M, Y)$.

The minimizers in the dual test are formally defined as:

$$\begin{aligned} \mu_{\text{dual}}^a(m, z, x) &:= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \int (y - \tilde{\mu}(m, z, x))^2 dP(Y^m = y, a, z, x) , \\ \mu_{\text{dual}}^a(m, x) &:= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \int (y - \tilde{\mu}(m, x))^2 dP(Y^m = y, a, x) . \end{aligned} \quad (76)$$

According to the identification of $P(X, Z, A, Y^m)$ in (74), the minimizers in (76) can be

re-expressed as weighted risk minimizers under \mathbb{P} via:

$$\begin{aligned}\mu_{\text{dual}}^a(m, z, x) &= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \mathbb{E} \left(\frac{1}{f_M(M | X, Z, A)} (Y - \tilde{\mu}(M, Z, X))^2 \right), \\ \mu_{\text{dual}}^a(m, x) &= \operatorname{argmin}_{\tilde{\mu} \in \mathcal{M}_\mu} \mathbb{E} \left(\frac{1}{f_M(M | X, Z, A)} (Y - \tilde{\mu}(M, X))^2 \right).\end{aligned}\tag{77}$$

A more stabilized version of the weighted risk minimizers in (77) can be obtained by replacing the inverse weight $1/f_M(M | X, Z, A)$ with the dual weight defined as

$$q_{\text{dual}}(M | A, Z, X) = f_M(M | a, Z, X) / f_M(M | A, Z, X),$$

motivated by the equivalence between the distribution $\mathbb{P}(X, Z, A, M^a, Y^a)$ (which is Markov relative to the graph in Fig. 4(c)) and the following weighted version of the observed distribution:

$$\begin{aligned}\mathbb{P}(X, Z, A, M^a, Y^a) &:= \mathbb{P}(X, Z, M^a, Y^a) \times q_{AY}(A | X, Z, M, Y) \\ &= q_{\text{dual}}(M | A, Z, X) \times \mathbb{P}(O).\end{aligned}$$

F.2 Details on a doubly robust test

F.2.1 Identification proofs

The parameters used in the test based on conditional counterfactual mean (CCM) are identified under three assumptions:

- (i) *Consistency*: $Y^m = Y$ when $M = m$, for all m in the state space of M .
- (ii) *Conditional ignorability*: $Y^m \perp M | A, Z, X$, for all m within its domain.
- (iii) *Positivity*: $p(M = m | A = a, Z = z, X = x) > 0$ for any (a, z, x) with $p(A = a, Z = z, X = x) > 0$, and $p(A = a | Z = z, X = x) > 0$ for any (z, x) with $p(Z = z, X = x) > 0$.

Given these identification assumptions, $\mu^m(z, x) := \mathbb{E}(Y^m | Z = z, X = x)$ is identified as:

$$\begin{aligned}\mu^m(z, x) &= \sum_a \mathbb{E}(Y^m | A = a, Z = z, X = x) p(A = a | Z = z, X = x) \\ &= \sum_a \mathbb{E}(Y | M = m, A = a, Z = z, X = x) p(A = a | Z = z, X = x),\end{aligned}$$

where the first equality follows from probability rules and the second equality follows from the consistency and conditional ignorability assumptions.

Given the identification of $\mu^m(z, x)$, the identification of $\mu^m(z)$ immediately follows by integrating $\mu^m(z, x)$ over the observed distribution of X .

F.2.2 Derivations of influence functions and one-step estimators

The one-step estimators of $\mu^m(z, x)$ and $\mu^m(z)$ can be obtained via deriving the corresponding influence functions.

Influence function for $\mu^m(z, x)$

$$\begin{aligned}
& \frac{\partial}{\partial \varepsilon} \mu^m(z, x) (\mathbf{P}_\varepsilon) \Big|_{\varepsilon=0} \\
&= \frac{\partial}{\partial \varepsilon} \int y' d\mathbf{P}_\varepsilon(y' | m, a', z, x) d\mathbf{P}_\varepsilon(a' | z, x) \Big|_{\varepsilon=0} \\
&= \int \frac{\mathbb{I}(m' = m, z' = z, x' = x)}{p(m | a', z, x) p(z, x)} (y' - \mathbb{E}(Y | m, a', z, x)) S(y' | m', a', z', x') d\mathbf{P}(y', m', a', z', x') \\
&\quad + \int \frac{\mathbb{I}(z' = z, x' = x)}{p(z, x)} (\mathbb{E}(Y | m, a', z, x) - \mu^m(z, x)) S(a' | z', x') d\mathbf{P}(a', z', x') \\
&= \int \frac{\mathbb{I}(m' = m, z' = z, x' = x)}{p(m | a', z, x) p(z, x)} (y' - \mathbb{E}(Y | m, a', z, x)) S(y', m', a', z', x') d\mathbf{P}(y', m', a', z', x') \\
&\quad + \int \frac{\mathbb{I}(z' = z, x' = x)}{p(z, x)} (\mathbb{E}(Y | m, a', z, x) - \mu^m(z, x)) S(a', z', x') d\mathbf{P}(a', z', x').
\end{aligned}$$

Given our notations, the np-EIF for $\mu^m(z, x)$ is given by:

$$\begin{aligned}
\Phi_{m,z,x}(\mathbf{Q})(O) &= \frac{\mathbb{I}(M = m, Z = z, X = x)}{f_M(m | A, z, x) p(z, x)} (Y - \mu(m, A, z, x)) \\
&\quad + \frac{\mathbb{I}(Z = z, X = x)}{p(z, x)} (\mu(m, A, z, x) - \mu^m(z, x)). \tag{78}
\end{aligned}$$

Influence function for $\mu^m(z)$

$$\begin{aligned}
& \frac{\partial}{\partial \varepsilon} \mu^m(z) (\mathbf{P}_\varepsilon) \Big|_{\varepsilon=0} \\
&= \frac{\partial}{\partial \varepsilon} \int y' d\mathbf{P}_\varepsilon(y' | m, a', z, x') d\mathbf{P}_\varepsilon(a' | z, x') d\mathbf{P}_\varepsilon(x') \Big|_{\varepsilon=0} \\
&= \int \frac{\mathbb{I}(m' = m, z' = z)}{p(m | a', z, x') p(z | x')} (y' - \mathbb{E}(Y | m, a', z, x')) S(y' | m', a', z', x') d\mathbf{P}(y', m', a', z', x') \\
&\quad + \int \frac{\mathbb{I}(z' = z)}{p(z | x')} (\mathbb{E}(Y | m, a', z, x') - \sum_{a^*} \mathbb{E}(Y | m, a^*, z, x') p(a^* | z, x')) S(a' | z', x') d\mathbf{P}(a', z', x')
\end{aligned}$$

$$+ \int \left(\sum_{a^*} \mathbb{E}(Y \mid m, a^*, z, x') p(a^* \mid z, x') - \mu^m(z) \right) S(x') dP(x') .$$

Given our notations, the np-EIF for $\mu^m(z)$ is given by:

$$\begin{aligned} \Phi_{m,z}(Q)(O) &= \frac{\mathbb{I}(M = m, Z = z)}{f_M(m \mid A, z, X) f_Z(z \mid X)} (Y - \mu(m, A, z, X)) \\ &\quad + \frac{\mathbb{I}(Z = z)}{f_Z(z \mid X)} (\mu(m, A, z, X) - \sum_a \mu(m, a, z, X) \pi(a \mid z, X)) \\ &\quad + \sum_a \mu(m, a, z, X) \pi(a \mid z, X) - \mu^m(z) . \end{aligned} \quad (79)$$

F.2.3 DR-CCM test for continuous X

Our proposed DR-CCM test is based on a TMLE for $\mu^m(z)$ in (35) that satisfies doubly robust asymptotic linearity. This ensures that both the test statistic and its confidence interval are consistently estimated if either $(\hat{\pi}, \hat{\mu})$ or (\hat{f}_M, \hat{f}_Z) is correctly specified. Achieving this property requires quantifying the first-order bias of the initial one-step estimator $\hat{\mu}^{+,m}(z)$ in (36) (and its TMLE counterpart), and updating the nuisance estimates to approximately solve both the score equation induced by the influence function of $\mu^m(z)$ and the estimating equations that render the first-order remainder bias negligible. Following the framework of [Van der Laan \[2014\]](#), [Benkeser \[2015\]](#), and [Benkeser et al. \[2017\]](#), we summarize the construction procedure below and refer readers to those works for further details.

Throughout this section, we assume that either $(\hat{\pi}, \hat{\mu})$ or (\hat{f}_M, \hat{f}_Z) is correctly specified, though we do not assume knowledge of which. Define $\xi(m, z, x; \mu) = \sum_a \pi(a \mid z, x) \mu(m, a, z, x)$. Given

the influence function $\Phi_{m,z}(Q)$ in (79), the R_2 term for $\mu^{+,m}(z)$ is derived as follows:

$$\begin{aligned}
R_2(\hat{Q}, Q) &= \hat{\mu}^{+,m}(z) - \mu^m(z) + \int \Phi_{m,z}(\hat{Q})(o) dP(o) \\
&= \int \left\{ \frac{f_M(m | a', z, x') \mathbb{I}(z' = z)}{\hat{f}_M(m | a', z, x') \hat{f}_Z(z | x')} (\mu(m, a', z, x') - \hat{\mu}(m, a', z, x')) \right. \\
&\quad + \frac{f_Z(z | x')}{\hat{f}_Z(z | x')} \sum_{a^*} \hat{\mu}(m, a^*, z, x') (\pi(a^* | z, x') - \hat{\pi}(a^* | z, x')) \\
&\quad \left. + \sum_{a^*} \hat{\mu}(m, a^*, z, x') \hat{\pi}(a^* | z, x') - \mu^m(z) \right\} dP(a', x') \\
&= \int \left\{ \frac{(f_M(m | a', z, x') - \hat{f}_M(m | a', z, x')) \mathbb{I}(z' = z)}{\hat{f}_M(m | a', z, x') \hat{f}_Z(z | x')} (\mu(m, a', z, x') - \hat{\mu}(m, a', z, x')) \right. \\
&\quad + \frac{f_Z(z | x') - \hat{f}_Z(z | x')}{\hat{f}_Z(z | x')} \sum_{a^*} \hat{\mu}(m, a^*, z, x') (\pi(a^* | z, x') - \hat{\pi}(a^* | z, x')) \\
&\quad \left. + \frac{f_Z(z | x') - \hat{f}_Z(z | x')}{\hat{f}_Z(z | x')} \sum_{a^*} \pi(a^* | z, x') (\mu(m, a^*, z, x') - \hat{\mu}(m, a^*, z, x')) \right\} dP(a', x'). \tag{80}
\end{aligned}$$

The R_2 term can be decomposed as $R_2(\hat{Q}, Q) = R_2^1(\hat{Q}, Q) + R_2^2(\hat{Q}, Q)$ where

$$\begin{aligned}
R_2^1(\hat{Q}, Q) &= \mathbb{E} \left(\frac{f_Z(z | X) - \hat{f}_Z(z | X)}{\hat{f}_Z(z | X)} (\xi(m, z, X; \mu) - \hat{\xi}(m, z, X; \hat{\mu})) \right), \text{ and} \\
R_2^2(\hat{Q}, Q) &= \mathbb{E} \left(\frac{\mathbb{I}(Z = z)}{\hat{f}_Z(z | X)} \frac{f_M(m | A, z, X) - \hat{f}_M(m | A, z, X)}{\hat{f}_M(m | A, z, X)} (\mu(m, A, z, X) - \hat{\mu}(m, A, z, X)) \right).
\end{aligned}$$

To analyze their behavior under model misspecification, let $(f_{M,+}, f_{Z,+})$ and (μ_+, ξ_+) denote the probability limits of the possibly misspecified nuisance estimates (\hat{f}_M, \hat{f}_Z) and $(\hat{\mu}, \hat{\xi})$, respectively. We further define the following two mapping functions:

$$\Phi_1(\tilde{\xi}) = \mathbb{E} \left(\frac{f_{Z,+} - f_Z}{f_{Z,+}} \tilde{\xi} \right), \quad \Gamma_0(\tilde{f}_Z) = \mathbb{E} \left(\frac{\xi_+ - \xi}{f_{Z,+}} \tilde{f}_Z \right).$$

These mappings can be used to describe the first-order behavior of $R_2^1(\hat{Q}, Q)$ as:

$$R_2^1(\hat{Q}, Q) = \{\Gamma_0(\hat{f}_Z) - \Gamma_0(f_Z)\} + \{\Phi_1(\hat{\xi}) - \Phi_1(\xi)\} + o_P(n^{-1/2}),$$

where the $o_P(n^{-1/2})$ term captures the second-order terms that are asymptotically negligible, given that the model for either f_Z or ξ is correctly specified.

A similar expansion applies to $R_2^2(\hat{Q}, Q)$. To derive it, we first define:

$$\Phi_2(\tilde{\mu}) = \mathbb{E} \left(\frac{\mathbb{I}(Z=z)}{\hat{f}_Z} \frac{f_{M,+} - f_M}{f_{M,+}} \tilde{\mu} \right), \quad \Gamma_1(\tilde{f}_M) = \mathbb{E} \left(\frac{\mathbb{I}(Z=z)}{\hat{f}_Z} \frac{\mu_+ - \mu}{f_M} \tilde{f}_M \right).$$

The first-order behavior of $R_2^2(\hat{Q}, Q)$ can be characterized as:

$$R_2^2(\hat{Q}, Q) = \{\Phi_2(\hat{\mu}) - \Phi_2(\mu_+)\} + \{\Gamma_1(\hat{f}_M) - \Gamma_1(f_M)\} + o_P(n^{-1/2}),$$

where $o_P(n^{-1/2})$ captures the second-order terms that are asymptotically negligible, provided that the model for either f_M or μ is correctly specified.

To achieve doubly-robust inference, the key lies in quantifying and correcting the first-order bias in the remainder term R_2 , defined in (80), using additional nuisance parameters that can be consistently estimated through nonparametric smoothing techniques at desired convergence rates. The general strategy for addressing model misspecification involves four main steps. We illustrate this below using the case where $(\mu_+, \xi_+) = (\mu, \xi)$ and $(f_{M,+}, f_{Z,+}) \neq (f_M, f_Z)$, in which case both $\Gamma_0(\hat{f}_Z) - \Gamma_0(f_Z)$ and $\Gamma_1(\hat{f}_M) - \Gamma_1(f_M)$ are zero:

1. Characterize the first-order behavior of each remainder term, $R_2^1 = \Phi_1(\hat{\xi}) - \Phi_1(\xi) + o_P(n^{-1/2})$,
 $R_2^2 = \Phi_2(\hat{\mu}) - \Phi_2(\mu) + o_P(n^{-1/2})$.
2. Approximate the first-order behavior of Φ_1 and Φ_2 by constructing mappings $\Phi_{1,n}$ and $\Phi_{2,n}$ that are estimable from the observed data. Specifically, we aim to represent $\Phi_1(\hat{\xi}) - \Phi_1(\xi) = \Phi_{1,n}(\hat{\xi}) - \Phi_{1,n}(\xi) + o_P(n^{-1/2})$, with an analogous expression holding for Φ_2 .
3. Perform linear expansions of $\Phi_{1,n}$ and $\Phi_{2,n}$ around ξ and μ , respectively. Taking $\Phi_{1,n}$ as an example, we have:

$$\Phi_{1,n}(\hat{\xi}) - \Phi_{1,n}(\xi) = P_n D_\xi(P) - P_n D_{\xi,n}(\hat{P}) + o_P(n^{-1/2}),$$

where $D_\xi(P)$ denotes the canonical gradient of $\Phi_{1,n}$ evaluated at the true distribution P , expressed in terms of several nuisance parameters to be defined later. The term $o_P(n^{-1/2})$ captures the empirical process term and second-order remainder term that are negligible. The empirical quantity $P_n D_{\xi,n}(\hat{P})$ represents the first-order bias of the R_2 term, which we

seek to correct.

4. Construct an estimator $\hat{\mu}^{\star,m}(z)$ that account for the first-order bias of each remainder term, such as $P_n D_{\xi,n}(\hat{P})$, thus establishing the asymptotic linearity.

Below, we illustrate Steps 1–4 using the case where $(\mu_+, \xi_+) \neq (\mu, \xi)$ and $(f_{M,+}, f_{Z,+}) = (f_M, f_Z)$, focusing on the expansion of R_2^1 . The approach for the complementary case, where $(\mu_+, \xi_+) = (\mu, \xi)$ and $(f_{M,+}, f_{Z,+}) \neq (f_M, f_Z)$, as well as for R_2^2 , follows analogously. A more detailed discussion of each scenario can be found in [Benkeser \[2015\]](#).

Under the discussed case, both $\Phi_1(\hat{\xi}) - \Phi_1(\xi)$ and $\Phi_2(\hat{\mu}) - \Phi_2(\mu)$ are zero, and we have

$$\begin{aligned}\Gamma_0(\hat{f}_Z) - \Gamma_0(f_Z) &= \mathbb{E}\left(\frac{\xi_+ - \xi}{\xi}(\hat{f}_Z - f_Z)\right) \\ &= -\mathbb{E}\left(\frac{\mathbb{I}(M = m, Z = z)}{f_M f_Z} \frac{Y - \xi_+}{f_Z} (\hat{f}_Z - f_Z)\right) \\ &= -\mathbb{E}\left(\frac{\xi^r}{f_Z^2}(\hat{f}_Z - f_Z)\right),\end{aligned}$$

where ξ^r is defined as:

$$\xi^r(X) = \mathbb{E}\left(\frac{\mathbb{I}(M = m, Z = z)}{f_M}(Y - \xi_+) \mid \hat{f}_Z, f_Z\right).$$

We note, however, that we cannot directly estimate $\xi^r(X)$ in practice because it involves unknown quantities. Thus, we proceed by approximating the first-order behavior of this quantity using a mapping that can be computed based only on the data. We proceed as

$$-\mathbb{E}\left(\frac{\xi^r}{f_Z^2}(\hat{f}_Z - f_Z)\right) = -\mathbb{E}\left(\frac{\hat{\xi}^r}{\hat{f}_Z^2}(\hat{f}_Z - f_Z)\right) + o_P(n^{-1/2}),$$

where $\hat{\xi}^r$ denotes an estimate of ξ^r , obtained by substituting the true nuisance parameters in its definition with their estimated counterparts and using nonparametric methods to estimate the conditional expectation through a univariate regression. These methods are assumed to yield consistent estimates at a sufficiently fast convergence rate. The term $o_P(n^{-1/2})$ captures second-order contributions that are negligible given that $(\mu_+, \xi_+) \neq (\mu, \xi)$ and $(f_{M,+}, f_{Z,+}) = (f_M, f_Z)$.

We conclude Step 2 with the following approximation:

$$\Gamma_0(\hat{f}_Z) - \Gamma_0(f_Z) = -\mathbb{I}(\mu_+ \neq \mu, \xi_+ \neq \xi, f_{M,+} = f_M, f_{Z,+} = f_Z) (\Gamma_{0,n}(\hat{f}_Z) - \Gamma_{0,n}(f_Z) + o_P(n^{-1/2})),$$

$$\text{where } \Gamma_{0,n}(\tilde{f}_Z) = \mathbb{E}\left(\frac{\hat{\xi}^r}{\hat{f}_Z^2} \tilde{f}_Z\right).$$

We then proceed to Step 3, where we perform a first-order expansion of $\Gamma_{0,n}$ around f_Z :

$$\Gamma_{0,n}(\hat{f}_Z) - \Gamma_{0,n}(f_Z) = (P_n - P) D_Z(\xi^r, f_Z) + P_n D_Z(\hat{\xi}^r, \hat{f}_Z) + o_P(n^{-1/2}),$$

where $D_Z(\xi^r, f_Z)$ is the canonical gradient of $\Gamma_{0,n}$, defined as $D_Z(\xi^r, f_Z) = \frac{\xi^r}{f_Z^2} (\mathbb{I}(Z = z) - f_Z)$, and $o_P(n^{-1/2})$ involves the empirical process term that is negligible.

This derivation isolates the first-order bias term $P_n D_Z(\hat{\xi}^r, \hat{f}_Z)$, which forms the basis for constructing the TMLE of interest. Recall that TMLE is a general framework for refining initial estimates of nuisance parameters through a targeting procedure. These updated nuisance estimates are designed to solve user-defined estimating equations, most commonly, the one associated with the efficient influence function of the estimand. To achieve doubly robust asymptotic linearity, the TMLE procedure is extended to solve additional estimating equations that correct for the first-order bias in the remainder term R_2 , such as ensuring that $P_n D_Z(\hat{\xi}^r, \hat{f}_Z)$ is negligible.

F.2.4 CCM test for discrete X

When X is discrete, we test pointwise invariance of $\mu^m(z, x)$ in z by defining $\Delta(m, x) := \mu^m(1, x) - \mu^m(0, x)$, where (see Appendix F.2.1)

$$\mu^m(z, x) = \sum_a \mu(m, a, z, x) \pi(a | z, x). \quad (81)$$

Let Δ denote the full collection of contrasts $\Delta(m, x)$ across all observed (m, x) pairs. Let Σ denote the variance-covariance matrix of Δ . Given estimates of $\mu^m(z, x)$, we obtain plug-in estimates Δ_n and Σ_n , and define the test statistic as $T_{n,CCM} := \Delta_n^\top \Sigma_n^{-1} \Delta_n$. Under the null, $T_{n,CCM}$ asymptotically follows a chi-squared distribution with d degrees of freedom, where d is the dimension of Δ (e.g., $d = 4$ when both M and X are binary). Alternatively, one may

perform univariate Wald tests for each (m, x) , adjusting for multiple comparisons via Bonferroni or Benjamini–Hochberg.

We next describe how to estimate $\mu^m(z, x)$, to ensure *doubly robust inference* for both (i) the *test statistic* and (ii) its *confidence intervals*.

Given that X, Z, M are discrete, we can achieve a root- n consistent estimator for $\mu^m(z, x)$ based on a simple plug-in estimate of (81). Alternatively, we can use a one-step estimator, which takes the following form, with corresponding EIF derived in Appendix F.2.2:

$$\hat{\mu}^{+,m}(z, x) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(Z_i = z, X_i = x)}{\hat{p}(z, x)} \left\{ \frac{\mathbb{I}(M_i = m)}{\hat{f}_M(m | A_i, z, x)} (Y_i - \hat{\mu}(m, A_i, z, x)) + \hat{\mu}(m, A_i, z, x) \right\}. \quad (82)$$

To construct a doubly robust confidence interval for $\hat{\mu}^{+,m}(z, x)$, we follow the approach in Appendix F.2.3, which requires the R_2 remainder term. This is derived below, using the EIF $\Phi_{m,z,x}(Q)$ given in (78).

$$\begin{aligned} R_2(\hat{Q}, Q) &:= \hat{\mu}^{+,m}(z, x) - \mu^m(z, x) + \int \Phi_{m,z,x}(\hat{Q})(o') dP(o') \\ &= \int \left\{ \frac{\mathbb{I}(z' = z, x' = x)}{\hat{p}(z, x)} \frac{f_M(m | a', z, x)}{\hat{f}_M(m | a', z, x)} (\mu(m, a', z, x) - \hat{\mu}(m, a', z, x)) \right\} dP(z', x', a') \\ &\quad + \frac{p(z, x)}{\hat{p}(z, x)} \sum_{a^*} \hat{\mu}(m, a^*, z, x) (\pi(a^* | z, x) - \hat{\pi}(a^* | z, x)) + \hat{\mu}^{+,m}(z, x) - \mu^m(z, x) \\ &= \int \frac{\mathbb{I}(z' = z, x' = x)}{\hat{p}(z, x)} \left\{ \sum_{a^*} \left\{ \frac{\pi(a^* | z, x)}{\hat{f}_M(m | a^*, z, x)} (f_M(m | a^*, z, x) - \hat{f}_M(m | a^*, z, x)) \right. \right. \\ &\quad \times (\mu(m, a^*, z, x) - \hat{\mu}(m, a^*, z, x)) \Big\} \Big\} dP(z', x') \\ &\quad + \frac{p(z, x) - \hat{p}(z, x)}{\hat{p}(z, x)} \sum_{a^*} \left\{ \hat{\mu}(m, a^*, z, x) (\pi(a^* | z, x) - \hat{\pi}(a^* | z, x)) \right. \\ &\quad \left. + \pi(a^* | z, x) (\mu(m, a^*, z, x) - \hat{\mu}(m, a^*, z, x)) \right\}. \end{aligned} \quad (83)$$

F.3 Details on efficiency gains under the Verma constraint

F.3.1 Identification proofs

Binary Z . To illustrate this, we first rewrite the ATE front-door functional in (1) to incorporate the anchor variable Z :

$$\psi(Q) = \iiint \sum_{a=0}^1 \mu(m, a, z, x) \pi(a | z, x) f_M(m | A = a_0, z, x) p(z, x) dm dz dx. \quad (84)$$

According to the Verma constraint in (73), the term $\sum_{a=0}^1 \mu(m, a, z, x) \pi(a | z, x)$ is invariant to the value z . Leveraging this invariance, we can fix z in the outcome and treatment models to a pre-specified level $z^* \in \mathcal{Z}$, which results in:

$$\psi_{z^*}(Q) = \iiint \sum_{a=0}^1 \mu(m, a, z^*, x) \pi(a | z^*, x) f_M(m | A = a_0, z, x) p(z, x) dm dz dx. \quad (85)$$

Continuous Z . When Z is continuous, we can construct a pathwise differentiable functional by leveraging the Verma constraint. Specifically, we equate $\sum_{a=0}^1 \mu(m, a, z^*, x) \pi(a | z^*, x)$ with the integral $\int \sum_{a=0}^1 \mu(m, a, z, x) p(a | z, x) \tilde{p}(z) dz$, where $\tilde{p}(Z)$ is a user-specified valid reference distribution for Z that need not match the true $p(Z)$. This yields the following identification formula:

$$\psi_{\tilde{p}}(Q) = \iiint \left\{ \int \sum_{a=0}^1 \mu(m, a, z, x) p(a | z, x) \tilde{p}(z) dz \right\} p(m | A = a_0, z, x) p(z, x) dm dz dx. \quad (86)$$

We denote this functional by $\psi_{\tilde{p}}$ to emphasize that the outcome regression and propensity score are integrated over the reference distribution $\tilde{p}(Z)$.

F.3.2 Nonparametric EIF derivation

Binary Z . The np-EIF for $\psi_{z^*}(Q)$ in (85) is derived as follows:

$$\begin{aligned} & \frac{\partial}{\partial \varepsilon} \psi_{z^*}(P_\varepsilon) \Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial \varepsilon} \int y p_\varepsilon(y | m, a, z^*, x) p_\varepsilon(m | a_0, z, x) p_\varepsilon(a | z^*, x) p_\varepsilon(z | x) p_\varepsilon(x) dy dm da dz dx \Big|_{\varepsilon=0} \end{aligned}$$

$$\begin{aligned}
&= \int \mathbb{I}(z = z^*) \frac{\sum_{z'} p(m | a_0, z', x) p(z' | x)}{p(m | a, z^*, x) p(z^* | x)} [y - \mathbb{E}(Y | m, a, z^*, x)] S(y | m, a, z, x) dP(o) \\
&\quad + \int \frac{\mathbb{I}(a = a_0)}{p(a | z, x)} [\xi_{z^*}(m, x) - \gamma_{z^*}(z, x)] S(m | a, z, x) dP(o) \\
&\quad + \int \frac{\mathbb{I}(z = z^*)}{p(z^* | x)} (a - \pi(1 | z^*, x)) \sum_{z'} [\kappa_{1,z^*}(z', x) - \kappa_{0,z^*}(z', x)] p(z' | x) S(z, x) dP(o) \\
&\quad + \int [\gamma_{z^*}(z, x) - \psi_{z^*}(\mathbf{Q})] S(z, x) dP(o) \\
&= \int \mathbb{I}(z = z^*) \frac{\sum_{z'} p(m | a_0, z', x) p(z' | x)}{p(m | a, z^*, x) p(z^* | x)} [y - \mathbb{E}(Y | m, a, z^*, x)] S(o) dP(o) \\
&\quad + \int \frac{\mathbb{I}(a = a_0)}{p(a | z, x)} [\xi_{z^*}(m, x) - \gamma_{z^*}(z, x)] S(o) dP(o) \\
&\quad + \int \frac{\mathbb{I}(z = z^*)}{p(z^* | x)} (a - \pi(1 | z^*, x)) \sum_{z'} [\kappa_{1,z^*}(z', x) - \kappa_{0,z^*}(z', x)] p(z' | x) S(z, x) dP(o) \\
&\quad + \int [\gamma_{z^*}(z, x) - \psi_{z^*}(\mathbf{Q})] S(z, x) dP(o).
\end{aligned}$$

Therefore, the np-EIF for $\psi_{z^*}(\mathbf{Q})$ is:

$$\begin{aligned}
\Phi_{z^*}(\mathbf{Q})(O_i) &= \frac{\mathbb{I}(Z_i = z^*)}{f_Z(z^* | X_i)} \sum_z f_{M,z^*}^r(M_i, A_i, z, X_i) f_Z(z | X_i) (Y_i - \mu(M_i, A_i, z^*, X_i)) \\
&\quad + \frac{\mathbb{I}(Z = z^*)}{f_Z(z^* | X_i)} (A_i - \pi(a | z^*, X_i)) \sum_z (\kappa_{1,z^*}(z, X_i) - \kappa_{0,z^*}(z, X_i)) f_Z(z | X_i) \\
&\quad + \frac{\mathbb{I}(A_i = a_0)}{\pi(a_0 | Z_i, X_i)} (\xi_{z^*}(M_i, X_i) - \gamma_{z^*}(Z_i, X_i)) + \gamma_{z^*}(Z_i, X_i) - \psi_{z^*}(\mathbf{Q}).
\end{aligned}$$

Continuous Z . The np-EIF for $\psi_{\tilde{p}}(\mathbf{Q})$ in (86) is derived as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \varepsilon} \psi_{\tilde{p}}(P_\varepsilon) \Big|_{\varepsilon=0} \\
&= \int \frac{\partial}{\partial \varepsilon} \psi_{z^*}(P_\varepsilon) \Big|_{\varepsilon=0} \tilde{p}(z^*) dz^* \\
&= \frac{\partial}{\partial \varepsilon} \int y \left[\int \sum_a p_\varepsilon(y | m, a, z^*, x) p_\varepsilon(a | z^*, x) \tilde{p}(z^*) dz^* \right] \\
&\quad \times p_\varepsilon(m | a_0, z, x) p_\varepsilon(z | x) p_\varepsilon(x) dy dm dz dx \Big|_{\varepsilon=0} \\
&= \int \tilde{p}(z^*) \frac{\int p(m | a_0, z, x) p(z | x) dz}{p(m | a, z^*, x) p(z^* | x)} [y - \mathbb{E}(Y | m, a, z^*, x)] S(y, m, a, z^*, x) P(y, m, a, z^*, x) \\
&\quad + \int \frac{\mathbb{I}(a = a_0)}{p(a | z, x)} [\xi_{\tilde{p}}(m, x) - \gamma_{\tilde{p}}(z, x)] S(m, a, z, x) dP(m, a, z, x) \\
&\quad + \int \frac{\tilde{p}(z^*)}{p(z^* | x)} (a - p(a = 1 | z^*, x)) \left[\int (\kappa_1(z, x) \right. \\
&\quad \left. - \kappa_0(z, x)) p(z | x) dz \right] S(a, z^*, x) dP(a, z^*, x)
\end{aligned}$$

$$+ \int [\gamma_{\tilde{p}}(z, x) - \psi_{\tilde{p}}(Q)] S(z, x) dP(z, x),$$

where $\xi_{\tilde{p}}(m, x) = \int \sum_a p(y | m, a, z^*, x) p(a | z^*, x) \tilde{p}(z^*) dz^*$, and $\gamma_{\tilde{p}}(z, x) = \int \xi_{\tilde{p}}(m, x) p(m | a_0, z, x) dm = \mathbb{E}(\xi_{\tilde{p}}(M, X) | a_0, z, x)$.

Let $Q = \{\mu, \gamma_{\tilde{p}}, \kappa_a, \pi, \xi_{\tilde{p}}, f_Z, p_{ZX}\}$. The np-EIF for $\psi_{\tilde{p}}(Q)$ is:

$$\begin{aligned} \Phi_{\tilde{p}}(Q)(O) &= \tilde{p}(Z) \frac{\int f_M(M | a_0, z, X) f_Z(z | X) dz}{f_M(M | a, Z, X) f_Z(Z | X)} [Y - \mu(M, A, Z, X)] \\ &\quad + \frac{\mathbb{I}(A = a_0)}{\pi(A | Z, X)} [\xi_{\tilde{p}}(M, X) - \gamma_{\tilde{p}}(Z, X)] \\ &\quad + \frac{\tilde{p}(Z)}{f_Z(Z | X)} (A - \pi(A = 1 | Z, X)) \int [\kappa_1(z, X) - \kappa_0(z, X)] f_Z(z | X) dz \\ &\quad + \gamma_{\tilde{p}}(Z, X) - \psi_{\tilde{p}}(P). \end{aligned}$$

Estimators of $\psi_{\tilde{p}}(Q)$ under univariate continuous Z . Constructing IF-based estimators now requires estimating the conditional densities f_M and f_Z . Given nuisance estimates \hat{Q} , the one-step estimator is given by:

$$\begin{aligned} \psi_{\tilde{p}}^+(\hat{Q}) &= \frac{1}{n} \sum_{i=1}^n \left[\tilde{p}(Z_i) \frac{\int \hat{f}_M(M_i | a_0, z, X_i) \hat{f}_Z(z | X_i) dz}{\hat{f}_M(M_i | a, Z_i, X_i) \hat{f}_Z(Z_i | X_i)} [Y - \hat{\mu}(M_i, A_i, Z_i, X_i)] \right. \\ &\quad + \frac{\mathbb{I}(A_i = a_0)}{\hat{\pi}(A_i | Z_i, X_i)} [\hat{\xi}_{\tilde{p}}(M_i, X_i) - \hat{\gamma}_{\tilde{p}}(Z_i, X_i)] \\ &\quad + \frac{\tilde{p}(Z_i)}{\hat{f}_Z(Z_i | X_i)} (A_i - \hat{\pi}(1 | Z_i, X_i)) \int [\hat{\kappa}_1(z, X_i) - \hat{\kappa}_0(z, X_i)] \hat{f}_Z(z | X_i) dz \\ &\quad \left. + \hat{\gamma}_{\tilde{p}}(Z_i, X_i) \right], \end{aligned}$$

where $\hat{\xi}_{\tilde{p}}(m, x)$ and $\hat{\gamma}_{\tilde{p}}(z, x)$ are nuisance estimates obtained via numerical integration with respect to the corresponding estimated conditional densities.

The Verma constraint enables the construction of a family of one-step estimators indexed by $\tilde{p}(Z)$. The choice of $\tilde{p}(Z)$ impacts the efficiency of the corresponding estimator. When Z is continuous, there are infinitely many valid choices of $\tilde{p}(Z)$ that respect the support of Z . As a result, identifying the optimal $\tilde{p}(Z)$, the one that minimizes asymptotic variance, becomes a more complex task. In such settings, a closed-form expression for the optimal $\tilde{p}(Z)$ is no longer available. Instead, we recommend that practitioners explore multiple choices of $\tilde{p}(Z)$, construct

the corresponding one-step estimators, and compare their estimated variances to guide selection.

F.3.3 Semiparametric gains: The optimal choice of α

The optimal weight α^{opt} aims to minimize the variance of $\psi_\alpha^+(Q)$, quantified by the IF as

$$\begin{aligned} & \mathbb{E}\left(\left\{\alpha\Phi_{z^*=1}(Q) + (1-\alpha)\Phi_{z^*=0}(Q)\right\}^2\right) \\ &= \alpha^2 \mathbb{E}(\Phi_{z^*=1}^2(Q)) + (1-\alpha)^2 \mathbb{E}(\Phi_{z^*=0}^2(Q)) + 2\alpha(1-\alpha)\mathbb{E}(\Phi_{z^*=1}(Q)\Phi_{z^*=0}(Q)). \end{aligned}$$

An optimizer is derived by differentiating the variance function with respect to α and setting the derivative to zero:

$$\begin{aligned} & \partial\mathbb{E}\left(\left\{\alpha\Phi_{z^*=1}(Q) + (1-\alpha)\Phi_{z^*=0}(Q)\right\}^2\right)/\partial\alpha \\ &= 2\alpha\mathbb{E}(\Phi_{z^*=1}^2(Q)) - 2(1-\alpha)\mathbb{E}(\Phi_{z^*=0}^2(Q)) + 2(1-2\alpha)\mathbb{E}(\Phi_{z^*=1}(Q)\Phi_{z^*=0}(Q)) = 0 \\ \implies & \alpha_{\text{opt}} = \mathbb{E}(\Phi_{z^*=0}(Q)(\Phi_{z^*=0}(Q) - \Phi_{z^*=1}(Q)))/\mathbb{E}((\Phi_{z^*=1}(Q) - \Phi_{z^*=0}(Q))^2). \end{aligned}$$

To prove α_{opt} minimizes the variance of $\psi^+\alpha(Q)$, we take the second derivative and show that it is greater than 0:

$$\begin{aligned} & \partial^2\mathbb{E}\left(\left\{\alpha\Phi_{z^*=1}(Q) + (1-\alpha)\Phi_{z^*=0}(Q)\right\}^2\right)/\partial\alpha^2 \\ &= 2\mathbb{E}(\Phi_{z^*=1}^2(Q)) + 2\mathbb{E}(\Phi_{z^*=0}^2(Q)) - 4\mathbb{E}(\Phi_{z^*=1}(Q)\Phi_{z^*=0}(Q)) \\ &= 2\mathbb{E}((\Phi_{z^*=1}(Q) - \Phi_{z^*=0}^2(Q))^2) \geq 0. \end{aligned}$$

G Details on simulation studies

G.1 Simulation 1: Theoretical properties

Summary. We evaluated the asymptotic properties of our ATE and ATT estimators established in Section 5. Specifically, we verified that (i) the \sqrt{n} -bias decayed at the expected rate, and (ii) the n -scaled variance converged to the efficient variance $P[\Phi(Q)^2]$. Simulations included univariate binary, univariate continuous, bivariate continuous, and four-dimensional continuous

mediators. Nuisance parameters were estimated using either fully parametric models or hybrid approaches combining parametric and kernel-based methods. Each scenario was replicated 1000 times at sample sizes ranging from 250 to 8000. Further details are provided below. Results (ATE: Figs (5)–(8); ATT: Figs (9)–(12)) confirmed that the estimators exhibited the expected large-sample behaviors. We also compared linear versus nonlinear (expit-based) submodels for the regression of the continuous outcome (see Appendix C.2) within the TMLE framework. Performance was evaluated in terms of bias, standard deviation (SD), mean squared error (MSE), 95% confidence interval (CI) coverage, and CI width. Comparisons were conducted for univariate binary, univariate continuous, and bivariate continuous mediators, across sample sizes of 500, 1,000, and 2,000. Additional details are provided below. Results (ATE: Table 5; ATT: Table 6) confirmed that both linear and nonlinear TMLEs yielded valid inference under correct model specification.

Detailed description of the DGPs used in Simulation 1 are provided below.

$$\begin{aligned}
X &\sim \text{Uniform}(0, 1), \\
A &\sim \text{Binomial}(0.3 + 0.2X), \\
U &\sim \text{Normal}(1 + A + X, 1), \\
(\text{univariate binary}) \quad M &\sim \text{Binomial}(\text{expit}(-1 + A + X)), \\
(\text{univariate continuous}) \quad M &\sim \text{Normal}(1 + A + X, 1), \\
(\text{bivariate continuous}) \quad M &\sim \text{Normal}\left(\begin{bmatrix} 1 + A + X \\ -1 - 0.5A + 2X \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right), \\
(\text{quadrivariate continuous}) \quad M &\sim \text{Normal}\left(\begin{bmatrix} 1 + A + X \\ -1 - 0.5A + 2X \\ -1 + 2A + X \\ 1 + 0.5A - X \end{bmatrix}, \begin{bmatrix} 5 & -1 & 0 & 2 \\ -1 & 6 & 1 & 0 \\ 0 & 1 & 4 & 3 \\ 2 & 0 & 3 & 7 \end{bmatrix}\right), \\
Y &\sim \text{Normal}(U + M + X, 1).
\end{aligned} \tag{87}$$

With *univariate binary* mediator, estimating the mediator density f_M through regressions is relatively straightforward. Consequently, $\psi_1(\hat{Q}^*)$, $\psi_1^+(\hat{Q})$, $\beta_1(\hat{Q}^*)$, and $\beta_1^+(\hat{Q})$ are identified

as the most suitable estimators. With *univariate continuous* mediator, we evaluate a total of ten estimators for ATE and ten estimators for ATT. In using estimators $\psi_1(\hat{Q}^*)$, $\psi_1^+(\hat{Q})$, $\beta_1(\hat{Q}^*)$, and $\beta_1^+(\hat{Q})$, we adopt the `np` package in R for a direct estimation of the mediator density. In using estimators $\psi_{2a}(\hat{Q}^*)$, $\psi_{2a}^+(\hat{Q})$, $\beta_a(\hat{Q}^*)$, and $\beta_a^+(\hat{Q})$, we adopt the `densratio` package for density ratio estimation. In using $\psi_{2b}(\hat{Q}^*)$, $\psi_{2b}^+(\hat{Q})$, $\beta_b(\hat{Q}^*)$, and $\beta_b^+(\hat{Q})$, we adopt the Bayes' rule for density ratio estimation. Additionally, we use modified versions of those sequential regression-based estimators, denoted by appending "`dnorm`" to their names. In these variants, we directly estimate the mediator density under the assumption that it follows a conditional Normal distribution. The `dnorm` estimators serve as benchmarks representing cases where the mediator density is correctly specified. With *multivariate mediators*, direct estimation of mediator densities can be challenging and computationally demanding. In applications, estimators that circumvent density estimation are preferred. Therefore, we only consider $\psi_{2a}(\hat{Q}^*)$, $\psi_{2a}^+(\hat{Q})$, $\beta_a(\hat{Q}^*)$, $\beta_a^+(\hat{Q})$, $\psi_{2b}(\hat{Q}^*)$, $\psi_{2b}^+(\hat{Q})$, $\beta_b(\hat{Q}^*)$, $\beta_b^+(\hat{Q})$, along with the variations where `dnorm` is used for mediator density ratio estimation, yielding a total of six estimators for both ATE and ATT estimation.

Figs (5)–(8) present the results establishing the \sqrt{n} -consistency of the proposed estimators for ATE, and Figs (9)–(12) are the corresponding results for ATT. In order, figures correspond to the settings with univariate binary, univariate continuous, bivariate continuous, and quadrivariate continuous mediators. In these figures, the left panel presents the \sqrt{n} -scaled bias and n -scaled variance as a function of sample size for the TMLE estimators, while the right panel presents results from the corresponding one-step estimators. The true variance in the variance plots is empirically calculated under the true DGP with a sample size of $n = 10^5$. Additionally, 95% confidence interval for each point estimate is derived and depicted as vertical bars in both the bias and variance plots. Sample standard deviation over 1000 multiple simulations is adopted for computing the confidence interval for each point estimate.

According to these figures, TMLE and one-step estimators are highly comparable under correct model specifications. We observe that estimators relying on nonparametric kernel density estimation or mediator density ratio estimation, as implemented via the `densratio` method, may face challenges in converging to the expected values. This issue is evident in both univariate and multivariate continuous mediator settings, even as the sample size grows. Overall, estimators

based on the Bayes' rule to estimate the density ratios are recommended due to their consistent performance in achieving the expected convergence results for most of the simulations.

We further compared TMLEs for the ATE and ATT using linear versus nonlinear submodels in the setting with a univariate continuous outcome. Linear submodels took the form $\tilde{\mu} = \hat{\mu} + \varepsilon_Y$, while nonlinear submodels followed the expit form detailed in (43) (Appendix C.2). Results for ATE and ATT are reported in Tables 5 and 6, respectively. Across submodel types, bias decreased with increasing sample size, and all estimators achieved nominal 95% CI coverage under correct nuisance specification—confirming the validity of both linear and nonlinear TMLEs.

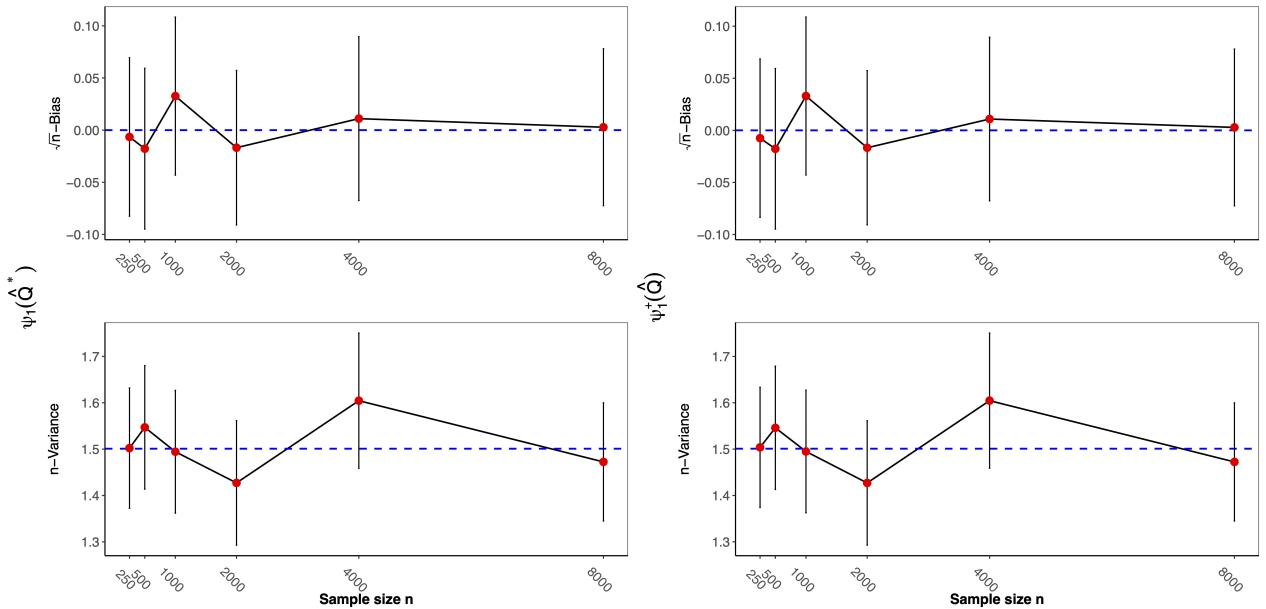


Figure 5: Simulation results validating the \sqrt{n} -consistency behaviors of the ATE estimators, under **univariate binary mediator**: (left) TMLE; (right) one-step estimator.

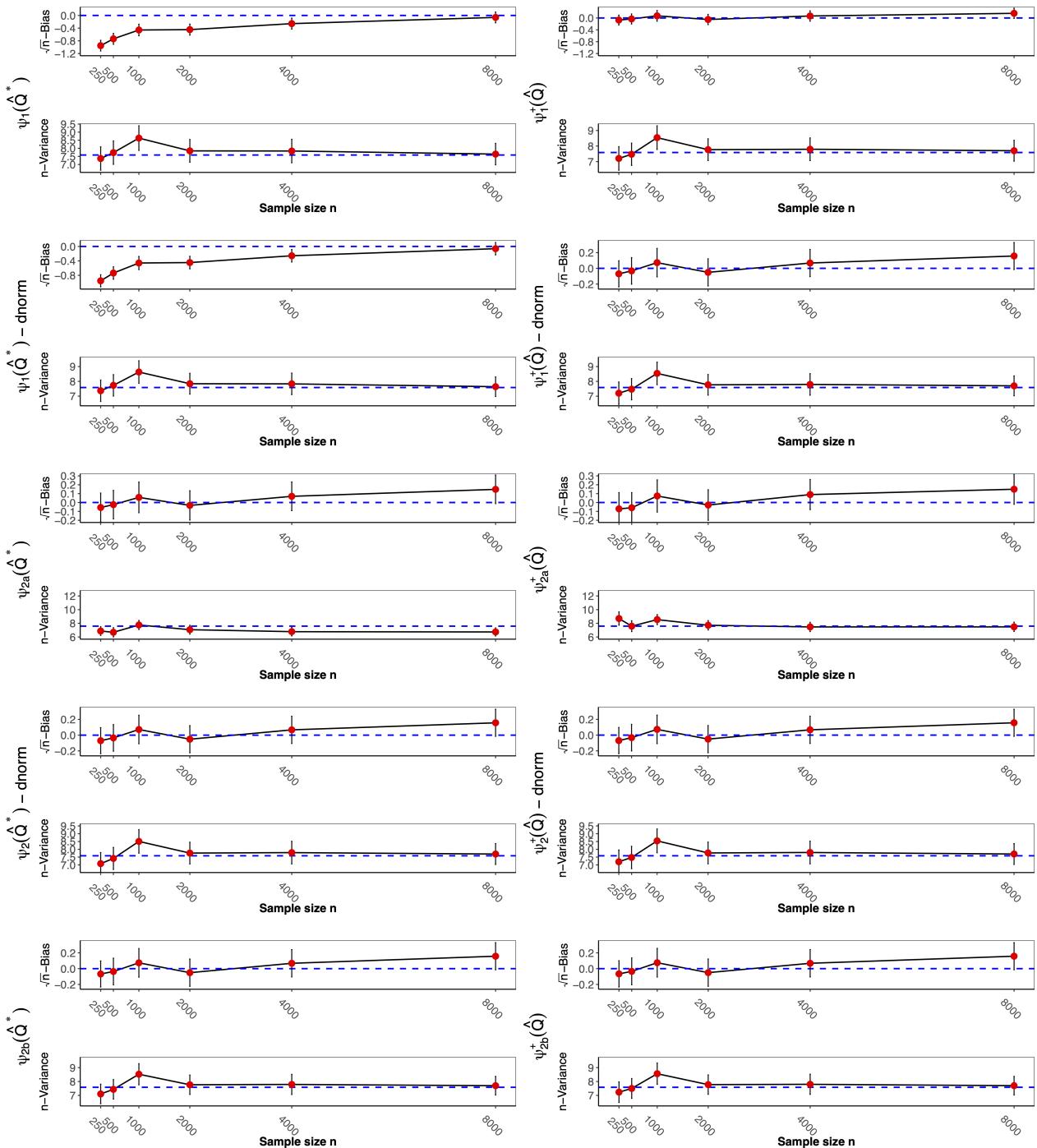


Figure 6: Simulation results validating the \sqrt{n} -consistency behaviors of the ATE estimators, under **univariate continuous mediator**: (left) TMLEs; (right) one-step estimators.

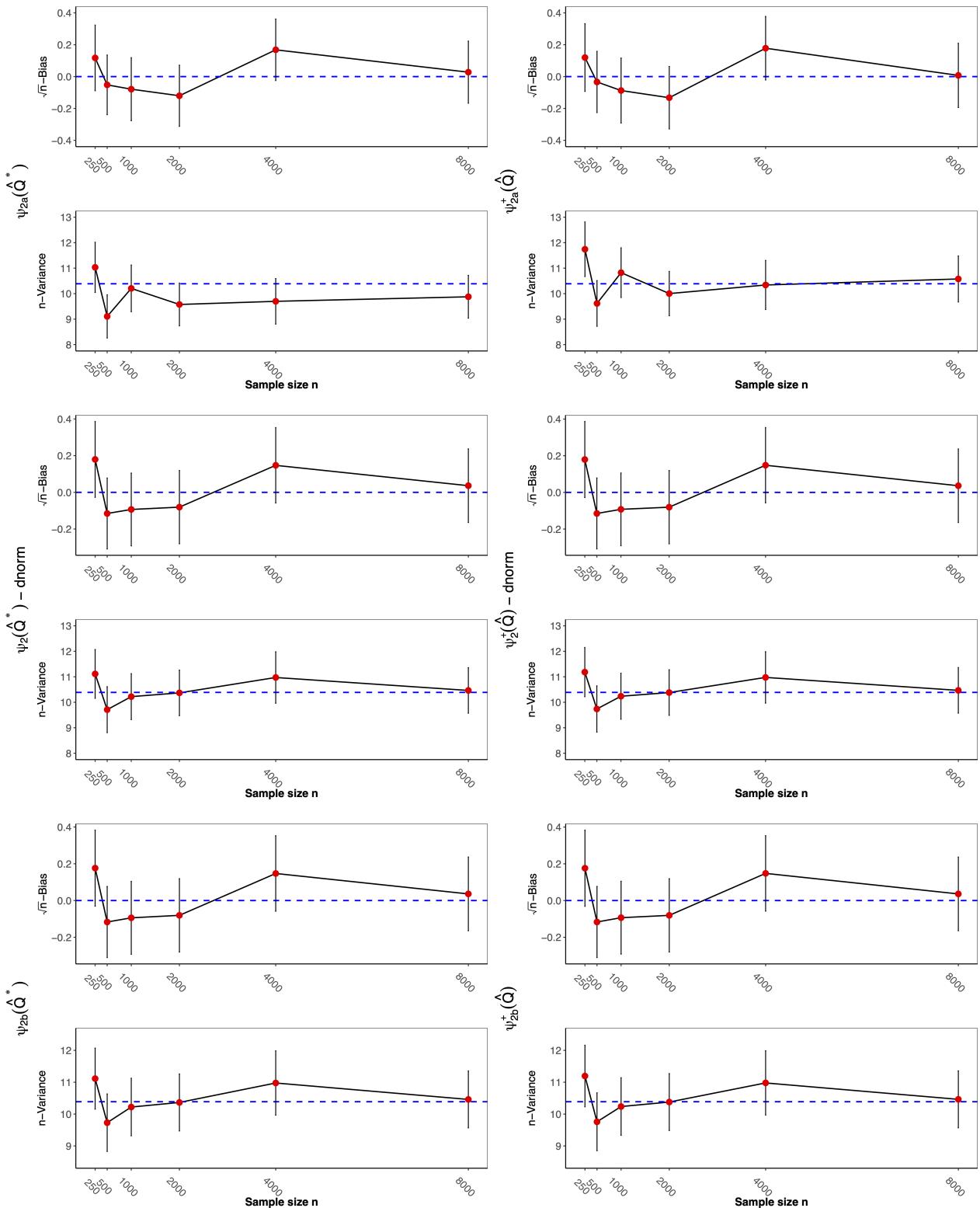


Figure 7: Simulation results validating the \sqrt{n} -consistency behaviors of the ATE estimators, under **bivariate continuous mediators**: (left) TMLEs; (right) one-step estimators.

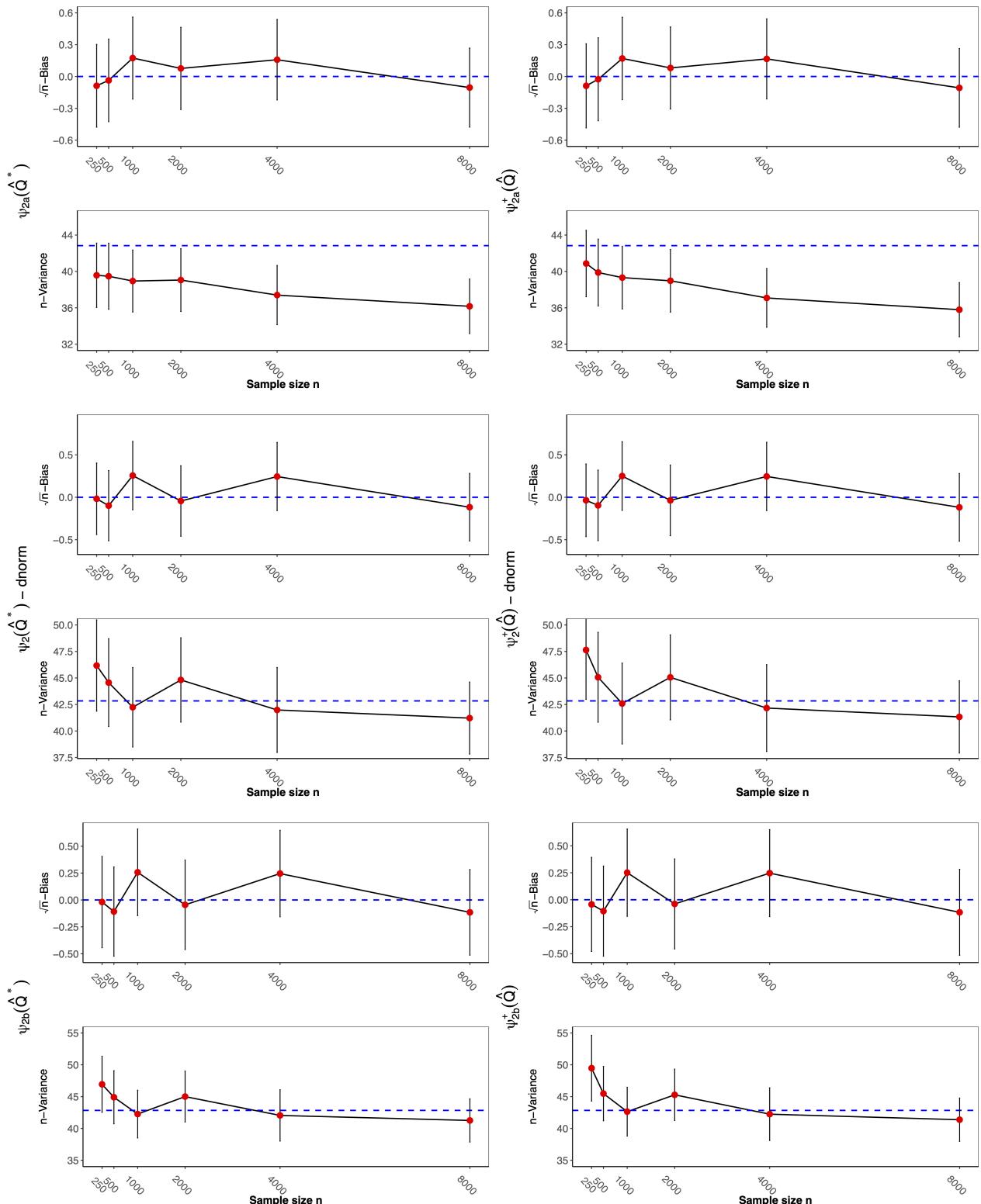


Figure 8: Simulation results validating the \sqrt{n} -consistency behaviors of the ATE estimators, under **quadrivariate continuous mediators**: (left) TMLEs; (right) one-step estimators.

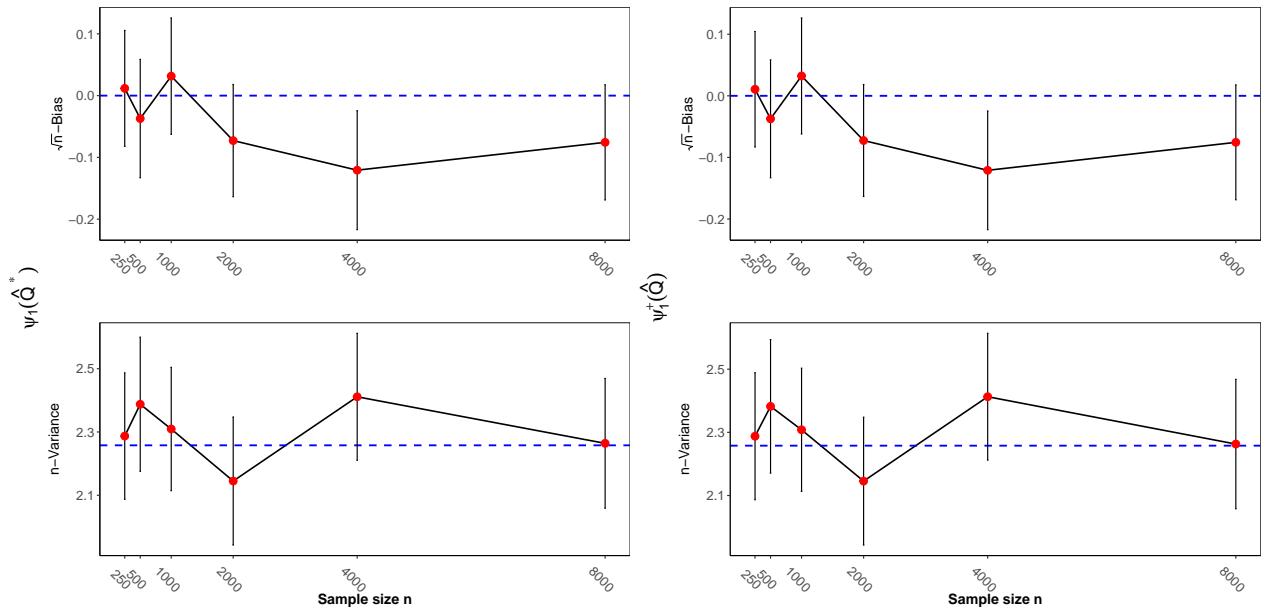


Figure 9: Simulation results validating the \sqrt{n} -consistency behaviors of the ATT estimators, under **univariate binary mediator**: (left) TMLEs; (right) one-step estimators.

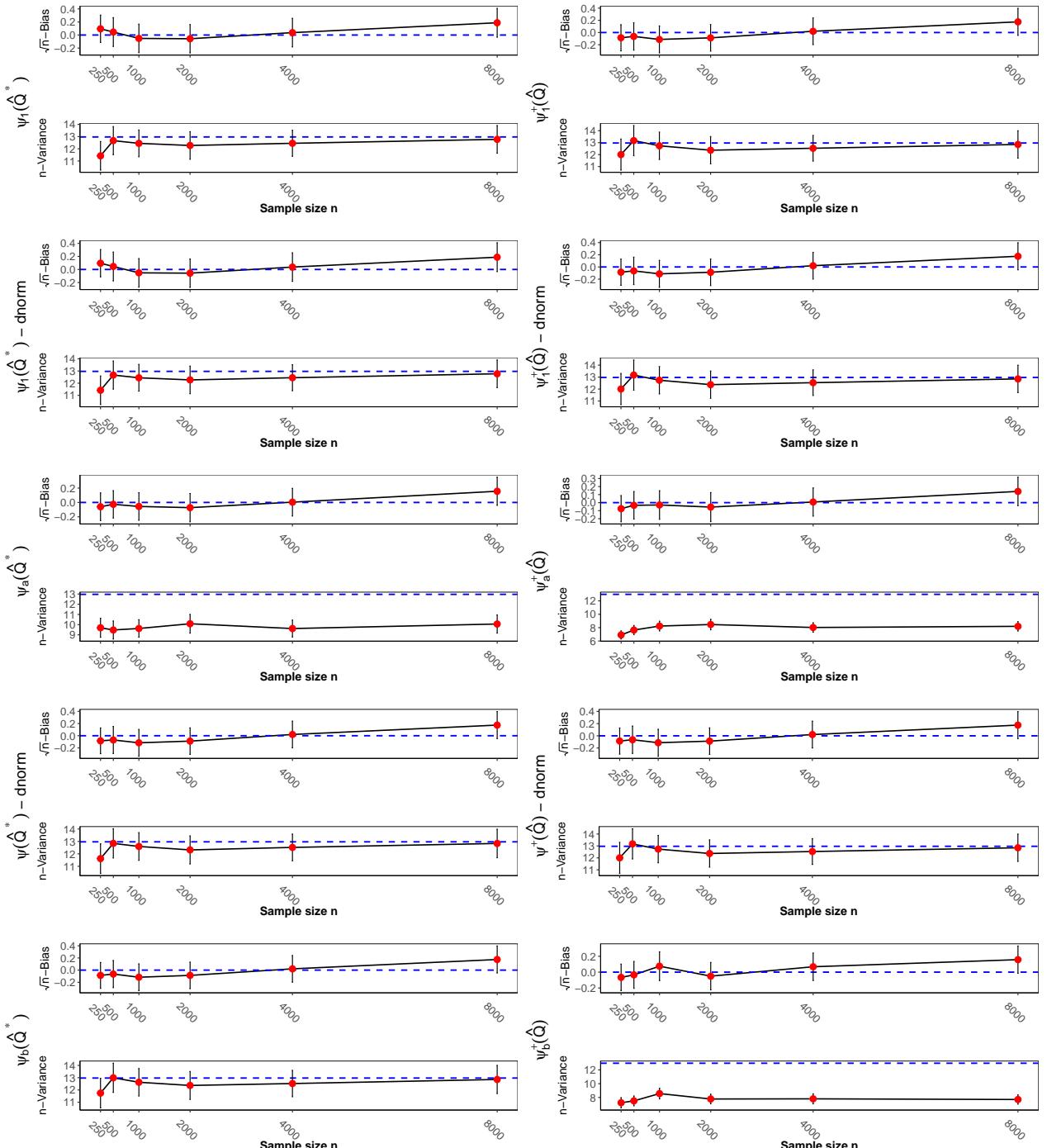


Figure 10: Simulation results validating the \sqrt{n} -consistency behaviors of the ATT estimators, under **univariate continuous mediator**: (left) TMLEs; (right) one-step estimators.

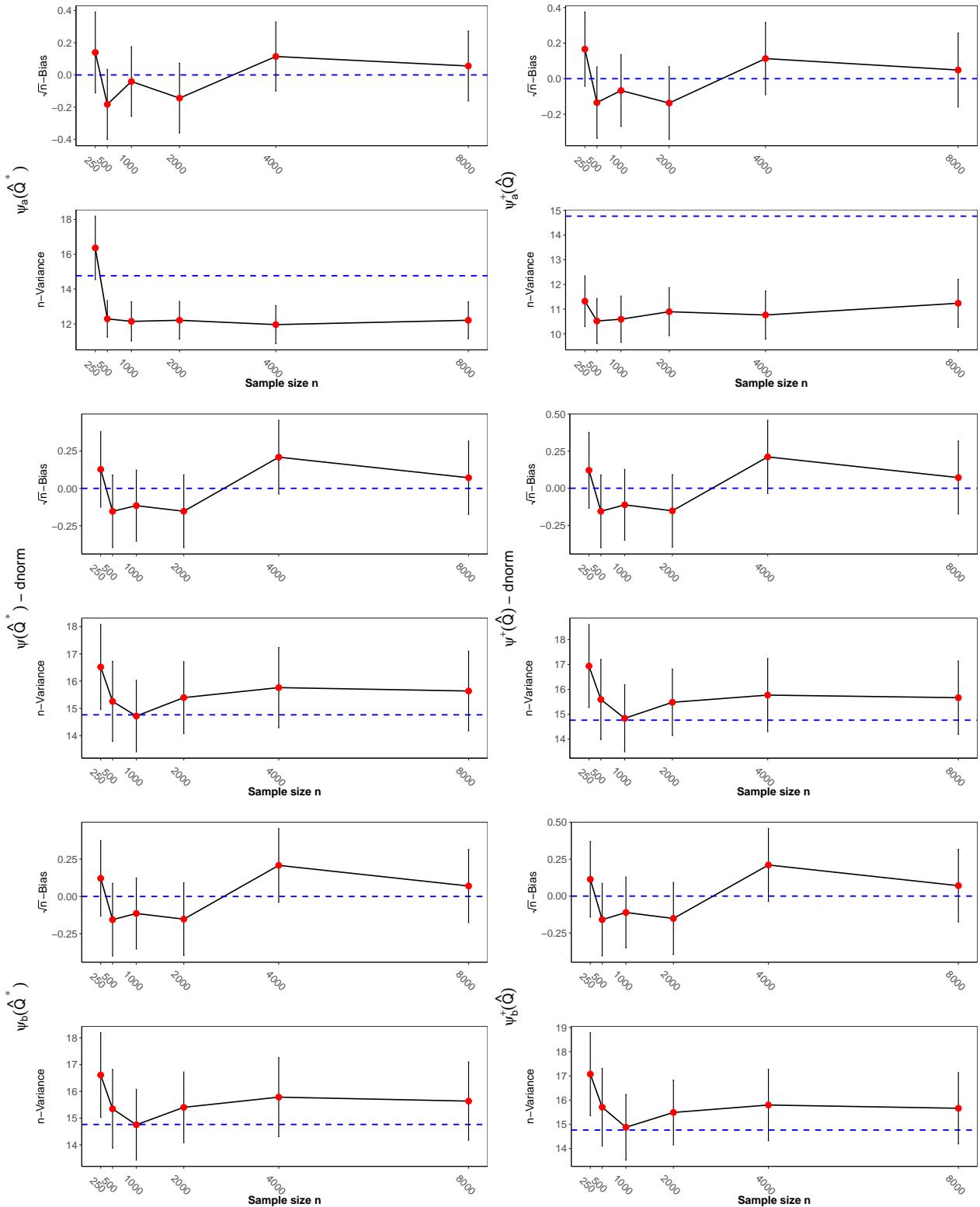


Figure 11: Simulation results validating the \sqrt{n} -consistency behaviors of the ATT estimators, under **bivariate continuous mediators**: : (left) TMLEs; (right) one-step estimators.

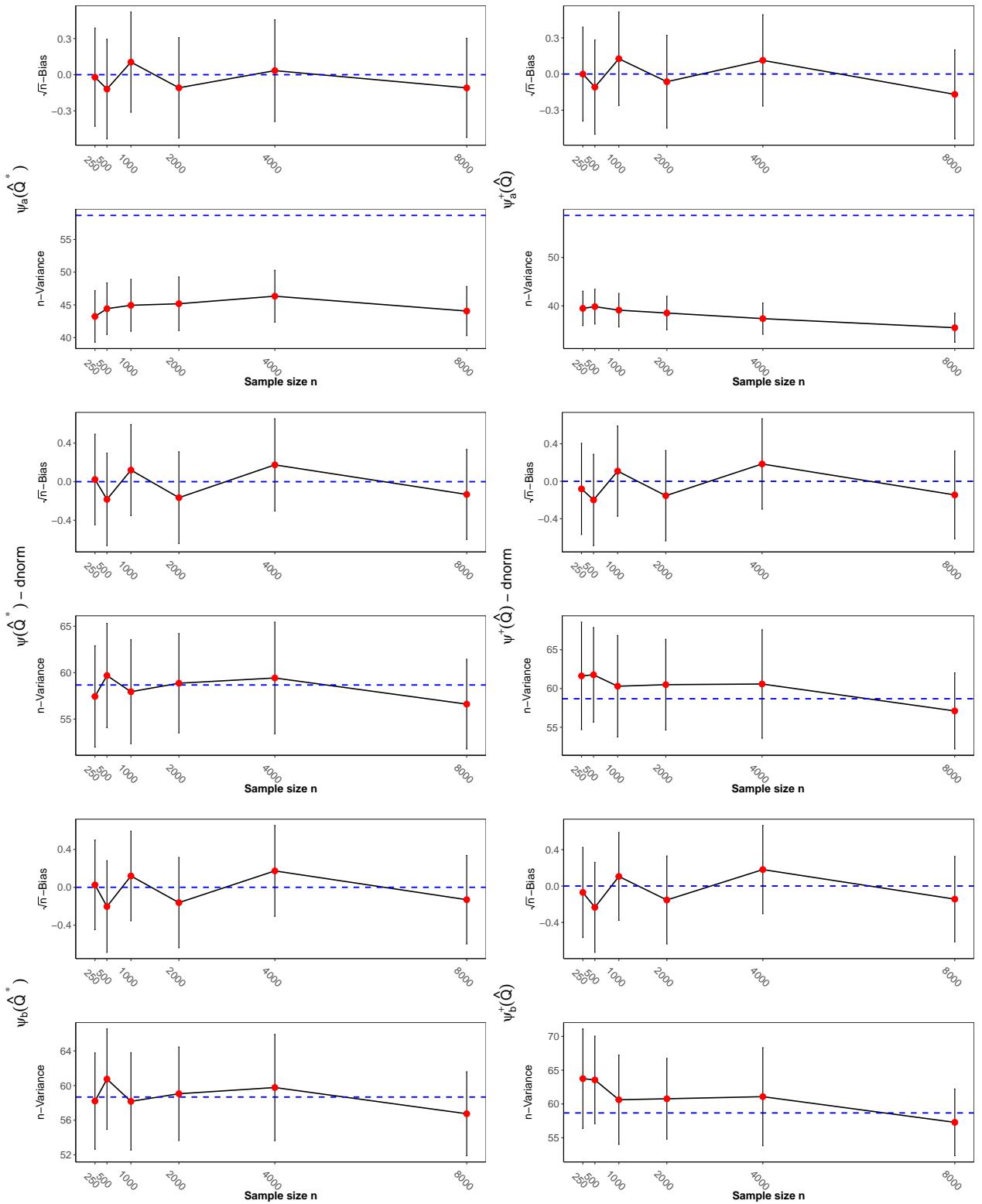


Figure 12: Simulation results validating the \sqrt{n} -consistency behaviors of the ATT estimators, under **quadrivariate continuous mediators**: : (left) TMLEs; (right) one-step estimators.

Table 5: Performance of ATE TMLEs under linear vs. expit outcome submodels across mediator types.

		Univariate Binary				Univariate Continuous				Bivariate Continuous					
		$\psi_1(\hat{Q}^*)$		$\psi_2(\hat{Q}^*) - dnorm$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_2(\hat{Q}^*) - dnorm$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$	
Submodels		Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit
n=500	Bias	-0.001	-0.001	-0.003	-0.005	-0.001	-0.002	-0.002	-0.002	-0.005	0.011	-0.002	-0.004	-0.005	-0.006
	SD	0.056	0.07	0.16	0.164	0.115	0.123	0.122	0.121	0.139	0.212	0.135	0.138	0.14	0.139
	MSE	0.003	0.005	0.026	0.027	0.013	0.015	0.015	0.015	0.019	0.022	0.018	0.019	0.02	0.019
	Coverage	94.1%	93.2%	92.7%	90.1%	96.3%	95.3%	94.9%	94.8%	95.2%	93.4%	95.7%	95.7%	95.4%	95.4%
n=1000	CI width	0.216	0.264	0.598	0.579	0.515	0.512	0.481	0.47	0.561	0.799	0.577	0.575	0.562	0.555
	Bias	0.001	0.002	-0.004	-0.005	0.002	0.002	0.002	0.002	-0.003	-0.005	-0.003	-0.003	-0.003	-0.003
	SD	0.039	0.048	0.112	0.113	0.088	0.089	0.092	0.091	0.101	0.14	0.101	0.1	0.101	0.101
	MSE	0.001	0.002	0.013	0.013	0.008	0.008	0.009	0.008	0.01	0.01	0.01	0.01	0.01	0.01
n=2000	Coverage	94.9%	94.6%	94.5%	92.9%	95.2%	95.2%	92.4%	92.4%	94.5%	95.3%	95%	95%	94.5%	94.2%
	CI width	0.152	0.187	0.433	0.425	0.361	0.362	0.342	0.337	0.396	0.563	0.409	0.408	0.396	0.394
	Bias	0	-0.001	-0.002	-0.002	-0.001	-0.001	-0.001	-0.001	-0.002	-0.003	-0.003	-0.003	-0.002	-0.002
	SD	0.027	0.033	0.079	0.079	0.06	0.06	0.062	0.062	0.072	0.101	0.069	0.069	0.072	0.072
	MSE	0.001	0.001	0.006	0.006	0.004	0.004	0.004	0.004	0.005	0.005	0.005	0.005	0.005	0.005
	Coverage	95.2%	95.1%	94.6%	93.9%	96.4%	96.6%	94.7%	94.5%	95.8%	94.5%	97.2%	97.2%	95.7%	95.5%
	CI width	0.107	0.132	0.308	0.304	0.257	0.257	0.241	0.239	0.281	0.396	0.291	0.291	0.281	0.28

Table 6: Performance of ATT TMLEs under linear vs. expit outcome submodels across mediator types.

		Univariate Binary				Univariate Continuous				Bivariate Continuous					
		$\psi_1(\hat{Q}^*)$		$\psi_2(\hat{Q}^*) - dnorm$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_2(\hat{Q}^*) - dnorm$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$	
Submodels		Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit	Linear	Logit
n=500	Bias	-0.002	-0.002	-0.003	-0.005	-0.001	-0.001	-0.003	-0.005	-0.007	-0.008	-0.008	-0.009	-0.007	-0.008
	SD	0.069	0.07	0.16	0.164	0.138	0.163	0.161	0.164	0.175	0.175	0.157	0.162	0.175	0.175
	MSE	0.005	0.005	0.026	0.027	0.019	0.026	0.026	0.027	0.031	0.031	0.025	0.026	0.031	0.031
	Coverage	93.5%	93.3%	92.7%	90.1%	90%	86.6%	92.5%	90.3%	94%	92.7%	91.3%	90.8%	93.9%	92.5%
n=1000	CI width	0.264	0.264	0.598	0.579	0.466	0.465	0.601	0.581	0.66	0.645	0.552	0.551	0.662	0.647
	Bias	0.001	0.001	-0.004	-0.005	-0.002	-0.003	-0.004	-0.005	-0.004	-0.004	-0.001	-0.001	-0.004	-0.004
	SD	0.048	0.048	0.112	0.113	0.098	0.104	0.112	0.113	0.121	0.121	0.11	0.11	0.122	0.121
	MSE	0.002	0.002	0.013	0.013	0.01	0.011	0.013	0.013	0.015	0.015	0.012	0.012	0.015	0.015
n=2000	Coverage	95%	94.7%	94.5%	92.9%	90%	87.6%	94.6%	93.1%	95%	94.1%	92.3%	92.7%	95%	94.2%
	CI width	0.187	0.187	0.433	0.425	0.331	0.33	0.433	0.426	0.467	0.462	0.39	0.39	0.468	0.462
	Bias	-0.002	-0.002	-0.002	-0.002	-0.002	-0.001	-0.002	-0.002	-0.003	-0.004	-0.003	-0.003	-0.003	-0.004
	SD	0.033	0.033	0.079	0.079	0.071	0.073	0.079	0.079	0.088	0.087	0.078	0.078	0.088	0.087
n=5000	MSE	0.001	0.001	0.006	0.006	0.005	0.005	0.006	0.006	0.008	0.008	0.006	0.006	0.008	0.008
	Coverage	95.4%	95.1%	94.6%	93.9%	91.4%	90%	94.6%	94%	94.3%	93.4%	92.2%	92.8%	94.4%	93.4%
n=10000	CI width	0.132	0.132	0.308	0.304	0.236	0.237	0.309	0.305	0.335	0.332	0.276	0.276	0.336	0.332

G.2 Simulation 2: Weak overlap

In this simulation, we compared the finite-sample characteristics of our proposed estimators for ATE and ATT in a setting with weak overlap. We generated the treatment variable according to Binomial ($0.001 + 0.998X$), while the rest of the DGPs, as specified in (87), remained unchanged.

Nuisance parameters were estimated as follows. Linear regressions and logistic regressions were employed to estimate $\mu(M, A, X)$ and $\pi(A | X)$, respectively. Logistic regression was utilized for estimating $f_M(M | A, X)$ under univariate binary mediator. For estimators $\psi_1(\hat{Q}^*)$, $\psi_1^+(\hat{Q})$,

Table 7: Comparison of ATT TMLE and one-step estimators under weak overlap across mediator types.

	Univariate Binary				Univariate Continuous				Bivariate Continuous			
	$\psi_1(\hat{Q}^*)$	$\psi_1^+(\hat{Q})$	$\psi_1(\hat{Q}^*)$	$\psi_1^+(\hat{Q})$	$\psi_{2a}(\hat{Q}^*)$	$\psi_{2a}^+(\hat{Q})$	$\psi_{2b}(\hat{Q}^*)$	$\psi_{2b}^+(\hat{Q})$	$\psi_{2a}(\hat{Q}^*)$	$\psi_{2a}^+(\hat{Q})$	$\psi_{2b}(\hat{Q}^*)$	$\psi_{2b}^+(\hat{Q})$
	Bias	SD	MSE	Coverage	Bias	SD	MSE	Coverage	Bias	SD	MSE	Coverage
$n=500$	-0.001	-0.01	0.006	0.031	0.004	0.045	0.01	0.021	-0.012	0.305	-0.023	-0.074
	0.103	0.613	0.161	1.147	0.415	3.644	0.398	1.748	0.5	5.278	0.456	2.021
	0.011	0.376	0.026	1.316	0.172	13.264	0.158	3.054	0.25	27.92	0.208	4.087
	87.4%	93.7%	97.6%	95.5%	97.2%	96.6%	97.2%	96.6%	98.9%	96.7%	97.2%	97.2%
$n=1000$	0.391	1.002	1.834	1.808	5.331	6.614	3.142	4.133	11.005	13.085	3.223	4.539
	0.004	0.011	0.002	-0.02	-0.002	0.108	-0.002	0.02	0.008	-0.004	0.007	0.03
	0.074	0.288	0.116	0.609	0.319	2.026	0.311	0.98	0.363	2.637	0.322	0.962
	0.006	0.083	0.013	0.37	0.102	4.113	0.097	0.959	0.131	6.946	0.104	0.926
$n=2000$	88.2%	94.5%	97.8%	95.6%	97.8%	97%	98%	97.1%	99.1%	97%	96.6%	95.3%
	0.299	0.555	1.055	1.042	3.444	3.975	2.072	2.476	5.411	6.096	2.058	2.443
	0.002	-0.001	0.002	0.013	0.007	-0.003	0.013	0.022	0.01	0.022	0.013	0.042
	0.051	0.17	0.08	0.343	0.24	1.035	0.244	0.64	0.304	1.463	0.282	0.678

$\beta_1(\hat{Q}^*)$, and $\beta_1^+(\hat{Q})$ in the case of a univariate continuous mediator, nonparametric kernel density estimation was applied to estimate $f_M(M \mid A, X)$ using the `np` package in R. For estimators $\psi_{2a}(\hat{Q}^*)$, $\psi_{2a}^+(\hat{Q})$, $\beta_a(\hat{Q}^*)$, and $\beta_a^+(\hat{Q})$ mediator density ratio was estimated via the `densratio` package in R. For estimators $\psi_{2b}(\hat{Q}^*)$, $\psi_{2b}^+(\hat{Q})$, $\beta_b(\hat{Q}^*)$, and $\beta_b^+(\hat{Q})$, the mediator density ratio was estimated using the reformulation presented in (16), where $\lambda(A \mid X, M)$ was estimated through logistic regressions.

Similar to Simulation 1, we evaluated the estimators based on bias, standard deviation (SD), mean squared error (MSE), 95% confidence interval (CI) coverage, and average 95% CI width. ATE estimation results are shown in Table 1 in the main manuscript. The ATT estimation results are provided in Table 7. Across all settings, TMLE and one-step estimators exhibited similar bias; however, TMLE typically achieved substantially lower SD, resulting in smaller overall MSE. This increased stability was also reflected in the CI width, which was generally narrower for TMLE, while maintaining comparable or more conservative coverage. These patterns held across both the smallest sample size ($n = 500$) and the largest ($n = 2000$).

G.3 Simulation 3: Model misspecification

Our third simulation explored the behavior of TMLEs and one-step estimators for both ATE and ATT in response to model misspecification, with a focus on univariate binary and univariate continuous mediators. We generated data as follows:

$$\begin{aligned} X &\sim \text{Uniform}(0, 1), & (\text{binary}) \quad M &\sim \text{Binomial}(\text{expit}(-1 + A + X - AX)), \\ A &\sim \text{Binomial}(\text{expit}(-1 + X)), & (\text{continuous}) \quad M &\sim \text{Normal}(1 + A + X - AX, 2), \\ U &\sim \text{Normal}(1 + A + X - AX, 2), & Y &\sim \text{Normal}(U + M + X - MX, 2). \end{aligned} \tag{88}$$

This simulation focused on quantifying the impact of nuisance parameter estimation on the final estimation of ATE and ATT. Comparisons between one-step and TMLE estimators were not the primary aim. Instead, we evaluated how a given estimator performs under inconsistent versus flexible estimation of Q. For the misspecified setting, we used main-effects linear regression models that excluded interaction terms present in the data-generating process. For flexible estimation, we employed the Super Learner algorithm [Van der Laan et al., 2007], an ensemble method that uses cross-validation to combine multiple candidate learners. These included intercept-only regression, generalized linear models (GLMs), Bayesian GLMs, multivariate adaptive regression splines, generalized additive models (GAMs), random forests, support vector machine (SVM), Bayesian Additive Regression Trees (BART), and extreme gradient boosting (XGBoost). Unlike the parametric models, these candidates are capable of capturing the interactions in the data. However, because many of them involve complex algorithms, they may violate the Donsker condition required by our theorems. To address this, we also implemented cross-fitted versions of each estimator.

We found that when misspecified working models were used for nuisance estimation, causal effect estimates were biased and CI coverage was poor across all sample sizes (see Table 2 in the main manuscript for ATE and Table 8 for ATT). In contrast, super learner-based estimators exhibited minimal bias across settings. CI coverage for these estimators generally improved with sample size, though some undercoverage was observed for the ψ_1 formulation of both the one-step and TMLE. These results suggest that for complex data-generating processes, flexible nuisance

Table 8: Performance of ATT estimators under model misspecifications across mediator types.

	TMLEs												One-step estimators													
	Univariate Binary						Univariate Continuous						Univariate Binary						Univariate Continuous							
	$\psi_1(\hat{Q}^*)$			$\psi_{2a}(\hat{Q}^*)$			$\psi_{2b}(\hat{Q}^*)$			$\psi_1^+(\hat{Q})$			$\psi_{2a}^+(\hat{Q})$			$\psi_{2b}^+(\hat{Q})$										
	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF	Linear	SL	CF		
n=500	Bias	-0.011	0.002	-0.006	-0.076	-0.011	-0.019	-0.077	-0.02	-0.028	-0.012	-0.001	-0.004	-0.076	-0.003	-0.006	-0.077	-0.019	-0.03							
	SD	0.055	0.063	0.062	0.1	0.143	0.145	0.1	0.11	0.125	0.056	0.06	0.067	0.1	0.113	0.113	0.1	0.11	0.128							
	MSE	0.003	0.004	0.004	0.016	0.02	0.021	0.016	0.013	0.016	0.003	0.004	0.005	0.016	0.013	0.013	0.016	0.013	0.017							
	Coverage	89.2%	88.3%	90.3%	88.1%	95.7%	95.2%	88%	92.5%	93.3%	90%	87.9%	89.3%	88.2%	98.1%	98.8%	88.1%	92.8%	93%							
	CI width	0.209	0.204	0.218	0.437	0.526	0.547	0.437	0.416	0.478	0.219	0.197	0.212	0.438	0.526	0.548	0.439	0.418	0.481							
	Bias	-0.011	0.001	-0.004	-0.077	-0.004	-0.007	-0.078	-0.014	-0.017	-0.012	-0.001	-0.003	-0.077	-0.002	-0.001	-0.078	-0.013	-0.017							
n=1000	SD	0.038	0.042	0.042	0.073	0.086	0.088	0.073	0.08	0.085	0.039	0.042	0.042	0.073	0.082	0.082	0.073	0.08	0.085							
	MSE	0.002	0.002	0.002	0.011	0.007	0.008	0.011	0.007	0.007	0.002	0.002	0.002	0.011	0.007	0.007	0.011	0.007	0.008							
	Coverage	89%	89.7%	89.6%	80.3%	96.4%	97%	80.3%	92.4%	92%	88.8%	89.2%	89.1%	80.4%	96.8%	98.1%	80.4%	92.8%	91.7%							
	CI width	0.145	0.147	0.151	0.308	0.366	0.374	0.308	0.293	0.316	0.152	0.144	0.148	0.309	0.366	0.375	0.309	0.294	0.317							
	Bias	-0.012	0.002	-0.001	-0.078	-0.004	-0.005	-0.078	-0.007	-0.01	-0.012	0.001	0	-0.078	-0.004	-0.004	-0.078	-0.007	-0.01							
	SD	0.026	0.028	0.029	0.051	0.059	0.06	0.051	0.055	0.058	0.027	0.028	0.029	0.051	0.056	0.056	0.051	0.055	0.058							
n=2000	MSE	0.001	0.001	0.001	0.009	0.003	0.004	0.009	0.003	0.003	0.001	0.001	0.001	0.009	0.003	0.003	0.009	0.003	0.003							
	Coverage	88.8%	93.5%	93.3%	67.8%	97.5%	97.2%	67.7%	93.7%	93.5%	90%	93.1%	93.5%	67.8%	97.7%	98.2%	67.7%	93.9%	93.6%							
	CI width	0.101	0.107	0.109	0.216	0.257	0.261	0.216	0.207	0.218	0.107	0.105	0.107	0.216	0.257	0.261	0.217	0.208	0.218							

estimation—such as super learner—is recommended to mitigate bias from model misspecification.

In this simulation, combining super learner with cross-fitting did not yield substantial gains in estimation performance.

G.4 Simulation 4: Cross-fitting

We examined the role of cross-fitting by focusing on random forests, which are known to perform poorly without sample splitting in high-dimensional settings [Chernozhukov et al., 2017, Biau, 2012]. We generated ten uniformly distributed confounders and introduced complex interactions and nonlinear terms between treatment, mediator, and covariates, as follows. Simulations used binary and continuous univariate mediators, with 1,000 replicates and sample sizes of 500, 1,000, and 2,000.

$$X_k \sim \text{Uniform}(0, 1), k \in \{1, \dots, 10\},$$

$$A \sim \text{Binomial}(\text{expit}(V_A [1 \ X \ X^2]^T)),$$

Binary mediator:

$$\begin{aligned} U &\sim \text{Normal} \left(V_U [1 \ A \ X \ AX_{1-5}]^T, 2 \right), \\ M &\sim \text{Binomial} \left(\text{expit} \left(V_M [1 \ A \ X \ AX_{1-5} \ X_{6-10}^2]^T \right) \right), \\ Y &\sim \text{Normal} \left(V_Y [U \ M \ X \ MX_{1-5} \ M^2 \ X_{6-10}^2]^T, 2 \right), \end{aligned} \tag{89}$$

Continuous mediator:

$$\begin{aligned} U &\sim \text{Normal} \left(V_U [1 \ A \ X \ AX_{1-5}]^T, 1 \right), \\ M &\sim \text{Normal} \left((V_M [1 \ A \ X \ AX_{1-5} \ X_{6-10}^2]^T), 1 \right), \\ Y &\sim \text{Normal} \left(V_Y [U \ M \ X \ MX_{1-5} \ M^2 \ X_{6-10}^2]^T, 1 \right), \end{aligned}$$

where

$$\begin{aligned} V_A &= 0.1 \times [0.48, 0.07, 1, -1, -0.34, -0.12, 0.3, -0.35, 1, -0.1, 0.46, 0.33, 0, \\ &\quad 0.45, 0.1, -0.32, -0.08, -0.2, 0.5, 0.5, -0.03], \\ V_U &= [-2, -1, -1, 2, 3, 0.5, 3, 2, -1, 1, -3, 1.5, -3, -2, 1, 3, 1.5], \\ V_M &= 0.025 \times [3, 1.5, -1.5, -1.5, -1, -2, -3, -3, -1.5, 2, 1.5, 3, 1.5, 2, 0.5, 0.5, 3, \\ &\quad -0.2, -0.33, 0.5, 0.3, -0.5], \\ V_Y &= [1, -2, -3, -1.5, 1, 0.5, -2, 1.5, -2, -3, -3, -1.5, -1, 0.5, 3, 1.5, 0.5, 3, 1, 1.5, -2, 3, -1] \\ X &= [X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}], \\ X_{1-5} &= [X_1, X_2, X_3, X_4, X_5], \\ X_{6-10} &= [X_6, X_7, X_8, X_9, X_{10}]. \end{aligned}$$

We implemented random forests using a standard set of tuning parameters: 500 trees were grown to a minimum node size of five observations for a continuous outcome and one observation for a binary variable. Cross-fitted ATE results are provided in Table 9. As shown in Table 9, cross-fitted ATE estimators consistently outperformed their non-cross-fitted counterparts, exhibiting lower bias and SD and substantially better CI coverage. Without cross-fitting, performance

degraded as sample size increased. These results underscore the importance of cross-fitting in high-dimensional or complex modeling settings. Cross-fitted ATT results are provided in Table 10.

We also repeated the simulation using a second set of tuning parameters. Specifically, we adopted a sparser random forest with 200 trees and a minimum node size of 1. Cross-fitted ATE and ATT results, under the sparser tuning parameter set, are provided in Table 11 and Table 12, respectively.

Tables 11 and 12 reveal a comparative analysis using a more sensitive random forest algorithm by increasing the variability of predictions. According to these results, the estimation performance of random forest is inferior, as evidenced by smaller CI coverage when compared with results produced by denser random forests (with 500 trees). In contrast, results yielded by performing sample splitting in conjunction with the sparser random forest remains highly comparable to those shown in Tables 9 and 10. These findings imply that in high-dimensional settings or scenarios where high estimation variance is anticipated from nuisance estimates, cross-fitting proves beneficial in reducing estimation bias and enhancing the stability of results.

G.5 Simulation 5: Model evaluation

Our fifth simulation evaluated the performance of proposed tests in scenarios that they are designed for. Performance was evaluated using type I error and power, which were calculated as the proportion of rejecting the null hypothesis during 200 simulation replicates for each test scenario. In each replicate, data were generated from a specific DGP, and the tests were applied. The rejection proportion corresponds to the type I error or power, depending on whether the DGP satisfies the front-door assumptions or not.

To evaluate type I error, we generated data from causal models that satisfy the front-door assumptions. We considered two model settings that differ in how Z relates to the other variables. In one setting, labeled as “DAG1”, Z has direct effects on both A and M . In another setting, labeled “DAG2”, Z has a direct effect on A and shares unmeasured confounding with M .

To evaluate power, we generated data from causal models that violate the front-door model assumptions. We considered two distinct settings, each representing a different type of violation. In both cases, Z had direct effects on both A and M . In a setting, labeled “DAG3”, violations

Table 9: Impact of cross-fitting on ATE TMLE and one-step estimators using random forests (RF: 500 trees; min node size = 5 for continuous, 1 for binary; CF: 5-fold cross-fitting).

		TMLEs						One-step estimators					
		Univariate Binary		Univariate Continuous		Univariate Binary		Univariate Continuous					
		$\psi_1(\hat{Q}^*)$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		$\psi_{2b}^+(\hat{Q})$	
		RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF
n=500	Bias	-0.162	-0.02	-0.312	0.055	-0.486	0.017	-0.103	-0.028	0.009	0.066	-0.492	0.014
	SD	0.166	0.14	0.372	0.331	0.369	0.285	0.051	0.128	0.432	0.318	0.373	0.286
	MSE	0.054	0.02	0.235	0.113	0.373	0.081	0.013	0.017	0.186	0.105	0.381	0.082
	Coverage	17.4%	82.8%	48.8%	86.9%	36.1%	87.3%	18.8%	86.3%	56.7%	87.6%	35.5%	87%
n=1000	CI width	0.128	0.389	0.681	0.98	0.717	0.862	0.119	0.388	0.682	0.977	0.718	0.861
	Bias	-0.162	-0.016	-0.329	0.054	-0.49	0.008	-0.1	-0.021	-0.017	0.059	-0.497	0.005
	SD	0.114	0.096	0.252	0.212	0.267	0.221	0.04	0.091	0.286	0.215	0.271	0.221
	MSE	0.039	0.009	0.172	0.048	0.312	0.049	0.012	0.009	0.082	0.049	0.32	0.049
n=2000	Coverage	13.3%	88.5%	30.1%	88.6%	19.5%	86.6%	12.4%	89.7%	52.4%	88.3%	18.3%	87.1%
	CI width	0.101	0.315	0.417	0.69	0.52	0.656	0.098	0.315	0.42	0.689	0.52	0.655
	Bias	-0.161	-0.01	-0.326	0.063	-0.473	0.019	-0.096	-0.013	-0.041	0.065	-0.479	0.016
	SD	0.083	0.074	0.176	0.148	0.186	0.164	0.034	0.072	0.197	0.15	0.189	0.164
	MSE	0.033	0.006	0.137	0.026	0.259	0.027	0.01	0.005	0.041	0.027	0.265	0.027
	Coverage	7.8%	90.4%	14.4%	89.8%	6.4%	86.5%	8.9%	90.7%	56.6%	88.9%	6.3%	86.5%
	CI width	0.081	0.246	0.292	0.52	0.376	0.499	0.08	0.246	0.294	0.519	0.376	0.499

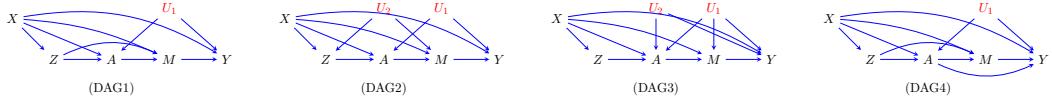


Figure 13: DAGs used in simulations on model evaluations: DAG1 and DAG2 correspond to scenarios where the front-door assumptions hold, while DAG3 and DAG4 depict scenarios where the assumptions are violated.

arose from unmeasured confounding between A and M , as well as between M and Y . In another setting, labeled ‘‘DAG4’’, the violation occurred through a direct effect of A on Y that was not mediated by M . The relationships among all variables in causal models DAG1-DAG4 are depicted in Fig. 13.

Under these four causal models, we designed three experimental settings to achieve different objectives by varying the types of variables (binary or continuous) for (X, Z, A, M, Y) and considering different forms of DGPs, including linear, quadratic, and interaction terms. The first setting aimed to verify that the proposed tests have approximately 0.05 type I error and show increasing power with larger sample sizes in the settings they were designed for. With the

Table 10: Impact of cross-fitting on ATT TMLE and one-step estimators using random forests (RF: 500 trees; min node size = 5 for continuous, 1 for binary; CF: 5-fold cross-fitting).

		TMLEs						One-step estimators					
		Univariate Binary		Univariate Continuous		Univariate Binary		Univariate Continuous		Univariate Binary		Univariate Continuous	
		$\psi_1(\hat{Q}^*)$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		$\psi_{2b}^+(\hat{Q})$	
		RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF
n=500	Bias	-0.513	-0.033	-0.1	0.311	-0.205	0.03	-0.471	-0.038	0.557	0.328	-0.21	0.057
	SD	0.26	0.142	0.238	0.429	0.196	0.311	0.208	0.133	0.384	0.376	0.148	0.326
	MSE	0.331	0.021	0.066	0.28	0.08	0.098	0.265	0.019	0.457	0.249	0.066	0.11
	Coverage	34.8%	84.5%	97.3%	71.1%	51.3%	86.3%	38%	86.1%	50.7%	72.1%	51.7%	84.5%
	CI width	0.828	0.413	1.108	1.047	0.441	0.923	0.827	0.409	1.115	1.045	0.441	0.926
n=1000	Bias	-0.507	-0.029	-0.148	0.161	-0.199	0.02	-0.463	-0.031	0.432	0.253	-0.211	0.042
	SD	0.186	0.1	0.148	0.237	0.143	0.239	0.151	0.097	0.247	0.259	0.11	0.247
	MSE	0.292	0.011	0.044	0.082	0.06	0.057	0.237	0.01	0.247	0.131	0.057	0.063
	Coverage	13.7%	88.7%	93.3%	78.4%	38.2%	84.7%	13.1%	89.4%	37.6%	65.7%	32.8%	83.3%
	CI width	0.59	0.329	0.722	0.709	0.323	0.692	0.591	0.327	0.726	0.709	0.322	0.693
n=2000	Bias	-0.508	-0.024	-0.129	0.172	-0.176	0.038	-0.462	-0.024	0.397	0.256	-0.2	0.057
	SD	0.129	0.075	0.097	0.161	0.097	0.168	0.107	0.074	0.153	0.177	0.076	0.172
	MSE	0.274	0.006	0.026	0.056	0.041	0.03	0.224	0.006	0.181	0.097	0.046	0.033
	Coverage	1.1%	89.6%	84.6%	70.6%	27.2%	87.4%	1.1%	90%	16%	48.5%	13%	85.5%
	CI width	0.421	0.254	0.454	0.528	0.235	0.52	0.422	0.253	0.456	0.528	0.235	0.521

second setting, we aimed to demonstrate the advantage of the DR-CCM test in providing valid inference even when some nuisance models are misspecified, by comparing its performance to other tests. The third setting aimed to showcase the flexibility of the proposed primal and dual tests in handling continuous mediators and incorporating machine learning models for nuisance estimation by examining scenarios with complex DGP forms, where both the anchor variable and the mediator are continuous.

In the first experimental setting, we evaluated the CCM, dual, and primal tests across three distinct variable-type configurations. In configuration 1, all variables, X , Z , A , M , and Y , were binary. In configuration 2, the outcome Y was continuous, while all other variables remained binary. In configuration 3, both X and Y were continuous, while Z , A , and M remained binary. Each test was evaluated at sample sizes of 500, 1000, 2000, 4000, and 10000.

With the first experimental setting, we found that all tests maintained type I error rates near the nominal level of 0.05 and exhibited increasing power with larger sample sizes. These results

Table 11: Impact of cross-fitting on ATE TMLE and one-step estimators using random forests (RF: 200 trees; min node size = 1; CF: 5-fold cross-fitting).

TMLEs												One-step estimators						
	Univariate Binary				Univariate Continuous				Univariate Binary				Univariate Continuous					
	$\psi_1(\hat{Q}^*)$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		$\psi_{2b}^+(\hat{Q})$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		$\psi_{2b}^+(\hat{Q})$	
	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF
n=500	Bias	-0.175	-0.021	-0.368	0.058	-0.518	0.02	-0.105	-0.027	-0.088	0.068	-0.524	0.018					
	SD	0.167	0.14	0.379	0.334	0.381	0.288	0.052	0.13	0.429	0.322	0.384	0.289					
	MSE	0.059	0.02	0.279	0.115	0.413	0.083	0.014	0.018	0.192	0.108	0.421	0.084					
	Coverage	18.1%	83.9%	42.4%	87%	33.9%	88.4%	19.6%	86.6%	53.9%	88.2%	33%	87.7%					
	CI width	0.133	0.397	0.657	0.995	0.743	0.88	0.122	0.396	0.657	0.992	0.744	0.879					
n=1000	Bias	-0.177	-0.016	-0.38	0.055	-0.52	0.013	-0.102	-0.02	-0.106	0.059	-0.525	0.01					
	SD	0.117	0.096	0.259	0.214	0.274	0.221	0.04	0.092	0.288	0.218	0.277	0.223					
	MSE	0.045	0.01	0.211	0.049	0.346	0.049	0.012	0.009	0.094	0.051	0.352	0.05					
	Coverage	11.7%	89.5%	23.2%	89%	17.4%	87.6%	12.6%	90.6%	49.3%	87.9%	17.5%	88.2%					
	CI width	0.105	0.32	0.412	0.7	0.535	0.666	0.101	0.32	0.414	0.699	0.535	0.666					
n=2000	Bias	-0.175	-0.01	-0.372	0.065	-0.498	0.025	-0.098	-0.012	-0.12	0.067	-0.504	0.021					
	SD	0.083	0.074	0.179	0.149	0.188	0.166	0.034	0.073	0.196	0.151	0.192	0.166					
	MSE	0.038	0.006	0.17	0.026	0.283	0.028	0.011	0.005	0.053	0.027	0.291	0.028					
	Coverage	5.7%	90.9%	9.9%	89.7%	4.9%	85.9%	8.5%	91.1%	50.9%	89.4%	5.2%	86.3%					
	CI width	0.084	0.25	0.294	0.526	0.384	0.506	0.082	0.25	0.295	0.525	0.385	0.505					

are summarized in Appendix Table 13.

Table 12: Impact of cross-fitting on ATT TMLE and one-step estimators using random forests (RF: 200 trees; min node size = 1; CF: 5-fold cross-fitting).

TMLEs												One-step estimators					
	Univariate Binary				Univariate Continuous				Univariate Binary				Univariate Continuous				
	$\psi_1(\hat{Q}^*)$		$\psi_{2a}(\hat{Q}^*)$		$\psi_{2b}(\hat{Q}^*)$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		$\psi_{2b}^+(\hat{Q})$		$\psi_1^+(\hat{Q})$		$\psi_{2a}^+(\hat{Q})$		
	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	RF	CF	
n=500	Bias	-0.406	-0.033	-0.123	0.312	-0.214	0.031	-0.114	-0.036	0.422	0.332	-0.217	0.062				
	SD	0.59	0.141	0.217	0.431	0.183	0.31	0.051	0.135	0.355	0.382	0.144	0.326				
	MSE	0.512	0.021	0.062	0.283	0.079	0.097	0.016	0.019	0.303	0.256	0.068	0.11				
	Coverage	4.7%	85.7%	96.6%	72.3%	46.6%	87.5%	21.8%	86.8%	59.3%	71.9%	47.5%	85.1%				
	CI width	0.275	0.424	1.01	1.073	0.408	0.956	0.147	0.421	1.018	1.072	0.409	0.959				
n=1000	Bias	-0.458	-0.03	-0.165	0.16	-0.207	0.019	-0.111	-0.031	0.314	0.253	-0.217	0.045				
	SD	0.442	0.1	0.137	0.237	0.134	0.241	0.041	0.098	0.226	0.261	0.106	0.251				
	MSE	0.405	0.011	0.046	0.082	0.061	0.059	0.014	0.011	0.15	0.132	0.058	0.065				
	Coverage	5.6%	90%	88.3%	79.6%	32.3%	87.1%	14.5%	90%	51.6%	66%	26.2%	83.9%				
	CI width	0.187	0.336	0.661	0.727	0.297	0.711	0.122	0.334	0.664	0.728	0.296	0.713				
n=2000	Bias	-0.49	-0.024	-0.146	0.171	-0.186	0.038	-0.106	-0.024	0.288	0.255	-0.206	0.059				
	SD	0.328	0.076	0.09	0.164	0.091	0.17	0.035	0.076	0.143	0.18	0.074	0.175				
	MSE	0.348	0.006	0.03	0.056	0.043	0.03	0.012	0.006	0.104	0.097	0.048	0.034				
	Coverage	4.3%	90.1%	76%	71.7%	18.1%	87.7%	9.3%	90.1%	25.5%	50.1%	10.1%	85.7%				
	CI width	0.136	0.26	0.416	0.539	0.215	0.532	0.1	0.258	0.418	0.539	0.215	0.533				

The DGPs for the **first experimental setting** with **variable-type configuration 1** (where all variables are binary) are displayed in (90).

(DAG1)

$$\begin{aligned}
U_1 &\sim \text{Binomial}(0.7), \\
X &\sim \text{Binomial}(0.3), \quad Z \sim \text{Binomial}(\text{expit}(-0.5 + 0.5X)) \\
A &\sim \text{Binomial}(\text{expit}(-0.5 - 1.1Z + 1.3U_1 + 0.5X + 1.75U_1Z - 1.2U_1X - 1.5ZX - 1.8U_1ZX)), \\
M &\sim \text{Binomial}(\text{expit}(-0.5 - A + 1.1Z - 0.5X - 1.25AZ + 1.5AX - 1.5ZX - 1.7AZX)), \\
Y &\sim \text{Binomial}(\text{expit}(-0.5 - 0.5M + U_1 + 0.5X - 1.2MU_1 + 1.5MX - 1.5U_1X - 1.7MU_1X)).
\end{aligned}$$

(DAG2)

$$\begin{aligned}
U_i &\sim \text{Binomial}(0.5), \quad i \in \{1, 2\}, \\
X &\sim \text{Binomial}(0.3), \quad Z \sim \text{Binomial}(\text{expit}(-0.5 + X + 1.5U_2 + 1.5XU_2)), \\
A &\sim \text{Binomial}(\text{expit}(-1 + Z + 1.5X + U_1 + 1.5ZX - 1.5U_1Z + 1.5U_1X - 1.7U_1ZX)), \\
M &\sim \text{Binomial}(\text{expit}(-1 + A + 1.5X + U_2 + 1.5AX - 1.5AU_2 + 1.5U_2X - 1.7AU_2X)), \\
Y &\sim \text{Binomial}(\text{expit}(-1 + 0.2M + 1.2X + U_1 + 1.5XU_1 - 1.5MX + 1.5MU_1 - 1.7MU_1X)).
\end{aligned}$$

(DAG3)

$$\begin{aligned}
U_i &\sim \text{Binomial}(0.5), \quad i \in \{1, 2\}, \\
X &\sim \text{Binomial}(0.5), \quad Z \sim \text{Binomial}(\text{expit}(-0.5 + 0.5X)), \\
A &\sim \text{Binomial}(\text{expit}(-0.5 + Z + 1.5X + U_1 + U_2 - 1.5ZX + 1.5ZU_1 - 1.5ZU_2 - 1.5XU_1 \\
&\quad + 1.5XU_2 - 1.5U_1U_2 - 1.7ZXU_1 + 1.2ZXU_2 - 1.7ZU_1U_2 - 1.7XU_1U_2 + 1.4ZXU_1U_2)), \\
M &\sim \text{Binomial}(\text{expit}(-1 + A + 1.5Z + X + U_1 - 1.5AZ + 1.5AX - 1.5AU_1 - 1.5ZX \\
&\quad + 1.5ZU_1 - 1.5XU_1 - 1.7AZX + 1.2AZU_1 - 1.7AXU_1 - 1.7ZXU_1 + 1.4AZXU_1)), \\
Y &\sim \text{Binomial}(\text{expit}(-0.5 + 0.5M + 0.2X + 1.2U_1 - 1.5U_2 - MX - 1.5MU_1 + MU_2 + 1.2XU_1 \\
&\quad + 0.5XU_2 + U_1U_2 + 1.1MXU_1 - 0.75MXU_2 - MU_1U_2 - 0.2XU_1U_2 + 0.5MXU_1U_2)).
\end{aligned}$$

(DAG4)

$$\begin{aligned}
P(U_1, X, Z, A, M) \text{ aligns with DAG1,} \\
Y &\sim \text{Binomial}(\text{expit}(-1 - 0.2M + 1.5A + 0.5X + 0.2U_1 - 1.2MA + 0.5MX + 0.3MU_1 - AX \\
&\quad + 0.5AU_1 - 0.5XU_1 + 0.5MAX - 0.5MAU_1 + 0.2MXU_1 - 0.5AXU_1 + MAXU_1)). \tag{90}
\end{aligned}$$

The DGPs for the **first experimental setting** with **variable-type configuration 2** (where Y is continuous, while all other variables are binary) are displayed in (91).

(DAG1)

$$\begin{aligned} U_1 &\sim \text{Binomial}(0.7), \\ X &\sim \text{Binomial}(0.3), \quad Z \sim \text{Binomial}(\text{expit}(-0.5 + 0.5X)), \\ A &\sim \text{Binomial}(\text{expit}(-0.5 - 1.1Z + 1.3U_1 + 0.5X)), \\ M &\sim \text{Binomial}(\text{expit}(-0.5 - A + 1.1Z - 0.5X)), \\ Y &\sim \text{Normal}(-0.5 - 0.5M + U_1 + 0.5X - 1.2MU_1, 0.5). \end{aligned}$$

(DAG2)

$$\begin{aligned} \text{P}(U_1, U_2, X, Z, A, M) &\text{ aligns with DAG2 in (90),} \\ Y &\sim \text{Normal}(-1 + 0.2M + 1.2X + U_1 + 1.5XU_1 - 1.5MX + 1.5MU_1 - 1.7MU_1X, 1). \end{aligned}$$

(DAG3)

$$\begin{aligned} \text{P}(U_1, U_2, X, Z, A, M) &\text{ aligns with DAG3 in (90),} \\ Y &\sim \text{Normal}(-0.5 + 0.5M + 0.2X + 1.2U_1 - 1.5U_2 - MX - 1.5MU_1 + MU_2 + 1.2XU_1 \\ &\quad + 0.5XU_2 + U_1U_2 + 1.1MXU_1 - 0.75MXU_2 - MU_1U_2 - 0.2XU_1U_2 + 0.5MXU_1U_2, 1). \end{aligned}$$

(DAG4)

$$\begin{aligned} \text{P}(U_1, U_2, X, Z) &\text{ aligns with DAG2 in (90),} \\ \text{P}(A, M \mid U, X, Z) &\text{ aligns with DAG1 in (90),} \\ Y &\sim \text{Normal}(-1 - 0.2M + 1.5A + 0.5X + 0.2U_1 - 1.2MA + 0.5MX + 0.3MU_1 - AX \\ &\quad + 0.5AU_1 - 0.5XU_1 + 0.5MAX - 0.5MAU_1 + 0.2MXU_1 - 0.5AXU_1 + MAXU_1, 1). \end{aligned} \tag{91}$$

The DGPs for the **first experimental setting** with **variable-type configuration 3** (where (X, Y) are continuous, while all other variables are binary) are displayed in (92).

(DAG1)

$P(U_1, Z, M)$ aligns with DAG1 in (91),

$$X \sim \text{Uniform}(0, 1), \quad A \sim \text{Binomial}((1 - 0.5Z + 1.3U_1 + 0.5X)/4),$$

$$Y \sim \text{Normal}(-0.5 - 0.5M + 0.5X, 1).$$

(DAG2)

$$U_i \sim \text{Binomial}(0.5), \quad i \in \{1, 2\},$$

$$X \sim \text{Uniform}(0, 1), \quad Z \sim \text{Binomial}((1 + X + 1.5U_2)/4),$$

$$A \sim \text{Binomial}((1 - 0.5Z + U_1 + 1.5X)/4), \quad M \sim \text{Binomial}(\text{expit}(-1 + A + 1.5X + U_2)),$$

$$Y \sim \text{Normal}(-1 + 0.2M + 1.2X, 1).$$

(DAG3)

(92)

$P(U_1, U_2, Z)$ aligns with DAG3 in (90),

$$X \sim \text{Uniform}(0, 1), \quad A \sim \text{Binomial}(\text{expit}(-0.5 + Z + U_1 + 1.5X)),$$

$$M \sim \text{Binomial}(\text{expit}(-1 + A + 1.5Z + X + U_1 - U_2)),$$

$$Y \sim \text{Normal}(-0.5 + 0.5M + 0.2X + 1.2U_1 - 1.5U_2, 0.5).$$

(DAG4)

$$U_1 \sim \text{Binomial}(0.7),$$

$$X \sim \mathcal{N}(1, 1), \quad A \sim \text{Binomial}(\text{expit}(-0.5 - 1.1Z + 1.3U_1 + 0.5X)),$$

$P(Z | X)$ aligns with DAG1 in (90),

$$M \sim \text{Binomial}(\text{expit}(-0.5 - A + 1.1Z - 0.5X)),$$

$$Y \sim \text{Normal}(-1 - 0.2M + 1.5A + 0.5X + 0.2U_1, 1).$$

In the second experimental setting, to demonstrate the advantage of the DR-CCM test, we compared its performance with that of the dual, and primal tests in a setting where the outcome regression could not be correctly specified using simple linear models. We considered a variable-type configuration in which both X and Y were continuous, while all other variables remained binary. Quadratic term X^2 and interaction term MX were added to the data-generating distribution of Y such that the outcome regression can no longer be correctly specified by simple linear models, creating condition of model misspecification to showcase the double robustness property of the DR-CCM test. As in previous evaluations, performance was assessed under the four causal models (DAG1–DAG4) across sample sizes of 500, 1000, 2000, 4000, and 10000.

In the second experimental setting, we observed that DR-CCM test was the only test among the four that consistently achieved type I error rates close to 0.05 while demonstrating increased power with larger sample sizes. In contrast, the other tests yielded increased type I error with larger sample sizes. These findings are presented in Table 3.

The DGPs for the **second experimental setting** are displayed in (93).

(DAG1)

$P(U_1, X, Z, A, M)$ aligns with DAG4 in (92),

$$Y \sim \text{Normal}(-0.5 - 0.5M + U_1 + 0.5X + 1.2X^2 - 1.5MX).$$

(DAG2)

$U_i \sim \text{Binomial}(0.5)$, $i \in \{1, 2\}$,

$X \sim \mathcal{N}(1, 0.5)$, $Z \sim \text{Binomial}(\text{expit}(-0.5 + X + 1.5U_2))$,

$A \sim \text{Binomial}(\text{expit}(-1 + A + 1.5X + U_1))$, $M \sim \text{Binomial}(\text{expit}(-1 + A + 1.5X + U_2))$,

$$Y \sim \text{Normal}(-1 + 0.2M + 1.2X + 1.2X^2 + 1.5MX + U_1, 0.5).$$

(DAG3)

$P(U_1, U_2, Z)$ aligns with DAG3 in (90),

$X \sim \mathcal{N}(1, 1)$, $A \sim \text{Binomial}(\text{expit}(-0.5 + Z + 1.5X + U_1))$,

$M \sim \text{Binomial}(\text{expit}(-1 + A + 1.5Z + X + 0.5X^2 + U_1 - U_2))$,

$$Y \sim \text{Normal}(-0.5 + 0.5M + 0.2X + 1.2X^2 + 1.2U_1 - 1.5U_2, 1).$$

(DAG4)

$U_1 \sim \text{Binomial}(0.5)$, $X \sim \mathcal{N}(1, 1)$,

$Z \sim \text{Binomial}(\text{expit}(-0.5 + 0.5X))$, $A \sim \text{Binomial}(\text{expit}(-0.5 - 1.1Z + 1.3U_1 + 0.5X))$,

$M \sim \text{Binomial}(\text{expit}(-0.5 - A + 1.1Z - 0.5X + 0.5X^2))$,

$$Y \sim \text{Normal}(-1 - 0.2M + 1.5A + 0.5X + 0.2U_1 + 0.5MX + 1.2X^2, 1).$$

(93)

In the third experimental setting, we further evaluated the performance of the dual and primal tests, with and without the use of flexible machine learning methods for model fitting, in a configuration where all variables except A were univariate continuous, and the outcome regression could not be correctly specified using simple linear models. When employing flexible methods, we used the Super Learner algorithm with two learners: the generalized linear model (`SL.glm`) and random forests (`SL.ranger`). This evaluation was conducted under three sample sizes: 500, 1000, and 2000.

With the third experimental setting, we found that incorporating Super Learners for nuisance model estimation helped keep type I error around 0.05 for both the dual and primal tests. In comparison, the tests without Super Learners had inflated type I errors, often exceeding 0.1 and increasing with sample sizes. While Super Learners helped keep type I error at the desired level, it came at the cost of reduced power relative to their non-Super Learner counterparts. These results are summarized in Table 14.

The DGPs for the **third experimental setting** are displayed in (94).

(DAG1)

$$U_1 \sim \text{Binomial}(0.7), \quad X \sim \text{Normal}(1, 1),$$

$$Z \sim \text{Normal}(-0.5 + 0.5X, 0.5), \quad A \sim \text{Binomial}(\text{expit}(-0.5 - 1.1Z + 0.5X)),$$

$$M \sim \text{Normal}(-0.5 - A + 1.1Z - 0.5X, 0.5), \quad Y \sim \text{Normal}(-0.5 - 0.5M + U_1 + 0.5X, 2).$$

(DAG2)

$$U_1 \sim \text{Binomial}(0.5), \quad U_2 \sim \mathcal{N}(1, 1),$$

$$X \sim \text{Normal}(1, 1), \quad Z \sim \text{Normal}(-0.5 + 1.5U_2, 0.5),$$

$$A \sim \text{Binomial}(\text{expit}(-1 + Z + 1.5X)), \quad M \sim \text{Normal}(-1 + A + 1.5X + U_2, 0.5),$$

$$Y \sim \text{Normal}(-1 + 0.2M + 1.2X + U_1, 1).$$

(DAG3)

$$U_i \sim \text{Binomial}(0.5), \quad i \in \{1, 2\},$$

$$X \sim \mathcal{N}(1, 1), \quad Z \sim \text{Normal}(-0.5 + 0.5X, 0.5),$$

$$A \sim \text{Binomial}(\text{expit}(-0.5 + Z + 1.5X + U_1 + U_2)), \quad M \sim \text{Normal}(-1 + A + 1.5Z + X + U_1, 0.5),$$

$$Y \sim \mathcal{N}(-0.5 + 0.5M + 0.2X + 1.2U_1 - 1.5U_2, 1).$$

(DAG4)

$$U_1 \sim \text{Binomial}(0.7), \quad X \sim \text{Uniform}(0.5, 1),$$

$$Z \sim \text{Uniform}(0, X), \quad A \sim \text{Binomial}((1 - 0.5Z + 1.3U_1 + 0.5X)/4),$$

$$M \sim \text{Normal}(-0.5 - A + 1.1Z - 0.5X, 0.2),$$

$$Y \sim \text{Normal}(-1 - 0.2M + 10A + 3AM - 0.5X + 0.2U_1, 2).$$

(94)

Table 13: Comparisons of the CCM, dual, and primal tests on type I error and power under model misspecification.

N	All binary					Y continuous				Y,X continuous			
	Type I error	Power		Type I error	Power		Type I error	Power		Type I error	Power		
	DAG1	DAG2	DAG3	DAG4	DAG1	DAG2	DAG3	DAG4	DAG1	DAG2	DAG3	DAG4	
CCM test													
500	0.07	0.12	0.215	0.07	0.06	0.225	0.73	0.15	0.06	0.05	0.245	0.69	
1000	0.03	0.06	0.455	0.06	0.035	0.16	0.98	0.33	0.06	0.065	0.395	0.935	
2000	0.035	0.035	0.87	0.07	0.015	0.125	1	0.565	0.04	0.08	0.64	1	
4000	0.02	0.055	0.995	0.195	0.035	0.05	1	0.93	0.05	0.055	0.93	1	
10000	0.045	0.025	1	0.69	0.05	0.05	1	1	0.05	0.06	1	1	
Dual test													
500	0.065	0.105	0.415	0.065	0.075	0.03	0.865	0.085	0.05	0.015	0.12	0.37	
1000	0.065	0.035	0.805	0.095	0.035	0.075	0.995	0.11	0.04	0.065	0.14	0.765	
2000	0.06	0.04	0.98	0.09	0.045	0.07	1	0.215	0.04	0.065	0.265	0.965	
4000	0.035	0.09	1	0.135	0.04	0.05	1	0.505	0.02	0.04	0.525	1	
10000	0.065	0.085	1	0.325	0.035	0.045	1	0.965	0.045	0.055	0.97	1	
Primal test													
500	0.07	0.04	0.37	0.04	0.07	0.06	0.285	0.08	0.055	0.035	0.09	0.17	
1000	0.025	0.01	0.73	0.055	0.055	0.025	0.515	0.05	0.07	0.06	0.09	0.22	
2000	0.02	0.025	0.985	0.05	0.055	0.04	0.855	0.11	0.06	0.08	0.15	0.495	
4000	0.025	0.015	1	0.13	0.055	0.02	0.995	0.1	0.04	0.045	0.345	0.65	
10000	0.02	0	1	0.57	0.06	0.04	1	0.275	0.08	0.035	0.86	0.94	

Table 14: Comparative analysis of dual and primal tests using linear vs Super Learners for complex DGPs.

Y,M,Z,X continuous									
N	Type I error				Power				
	DAG1		DAG2		DAG3		DAG4		
Nuisance models	Linear	SL	Linear	SL	Linear	SL	Linear	SL	
Dual test									
500	0.065	0.045	0.11	0.055	0.465	0.05	0.35	0.07	
1000	0.055	0.06	0.155	0.05	0.685	0.08	0.595	0.105	
2000	0.135	0.05	0.175	0.055	0.84	0.11	0.725	0.145	
Primal test									
500	0.165	0.05	0.13	0.055	0.915	0.115	0.555	0.445	
1000	0.18	0.06	0.145	0.08	0.935	0.165	0.5	0.485	
2000	0.185	0.035	0.155	0.06	0.975	0.205	0.39	0.42	

G.6 Simulation 6: Efficiency gain

This simulation investigated the efficiency gains from leveraging the Verma constraint, considering scenarios where Z is either univariate binary or continuous. The DGP for binary Z is given in (95), and the DGP for continuous Z in (96).

$$\begin{aligned}
 U &\sim \text{Normal}(1 + A - 0.2Z, 1), \\
 Z &\sim \text{Binomial}(0.2), \quad A \sim \text{Binomial}(0.3 + 0.2Z), \\
 M &\sim \text{Binomial}(\text{expit}(-1 + A + Z)), \quad Y \sim \text{Normal}(U + M, 1).
 \end{aligned} \tag{95}$$

$$\begin{aligned}
U &\sim \text{Normal}(1 + A - \text{expit}(0.3 + 0.2Z), 0.1) \\
Z &\sim \text{Normal}(1, 1), \quad A \sim \text{Binomial}(\text{expit}(0.3 + 0.2Z)) \\
M &\sim \text{Binomial}(\text{expit}(-1 + A + Z)), \quad Y \sim \text{Normal}(U + M, 0.1).
\end{aligned} \tag{96}$$

For binary Z , results are shown in Fig. 14. The estimator $\psi_{\text{opt}}^+(\hat{Q})$ exhibited lower asymptotic variance than both $\psi_{z^*=1}^+(\hat{Q})$ and $\psi_{z^*=0}^+(\hat{Q})$, reducing variance by half compared to $\psi_{z^*=1}^+(\hat{Q})$. Notably, in our simulations, $\psi_{\text{opt}}^+(\hat{Q})$ achieved the same variance as an estimator using weight $\alpha = \hat{P}(Z = 1)$, the marginal distribution of Z .

For continuous Z , results are shown in Fig. 15. Among the three choices of $\tilde{p}(Z)$, using the density of a $\text{Normal}(10, 1)$ distribution yielded the lowest variance, even outperforming the true data-generating density $P(Z)$, which had the second lowest variance.

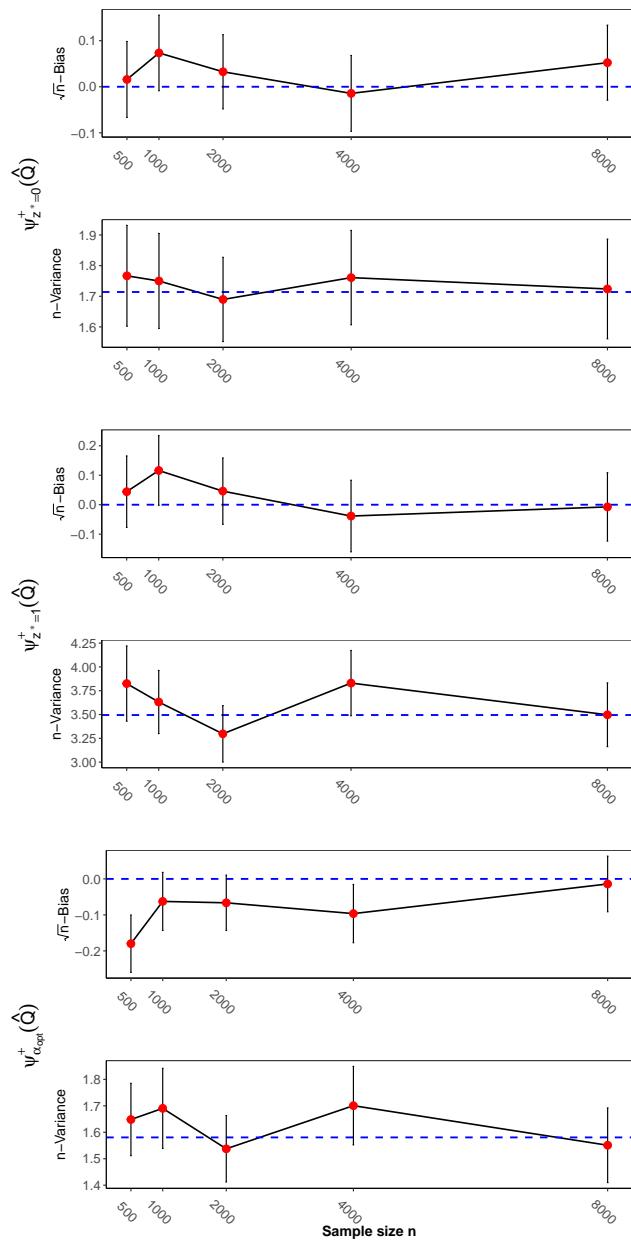


Figure 14: Simulation results demonstrating efficiency gains in ATE estimation when utilizing the Verma constraint under binary Z .

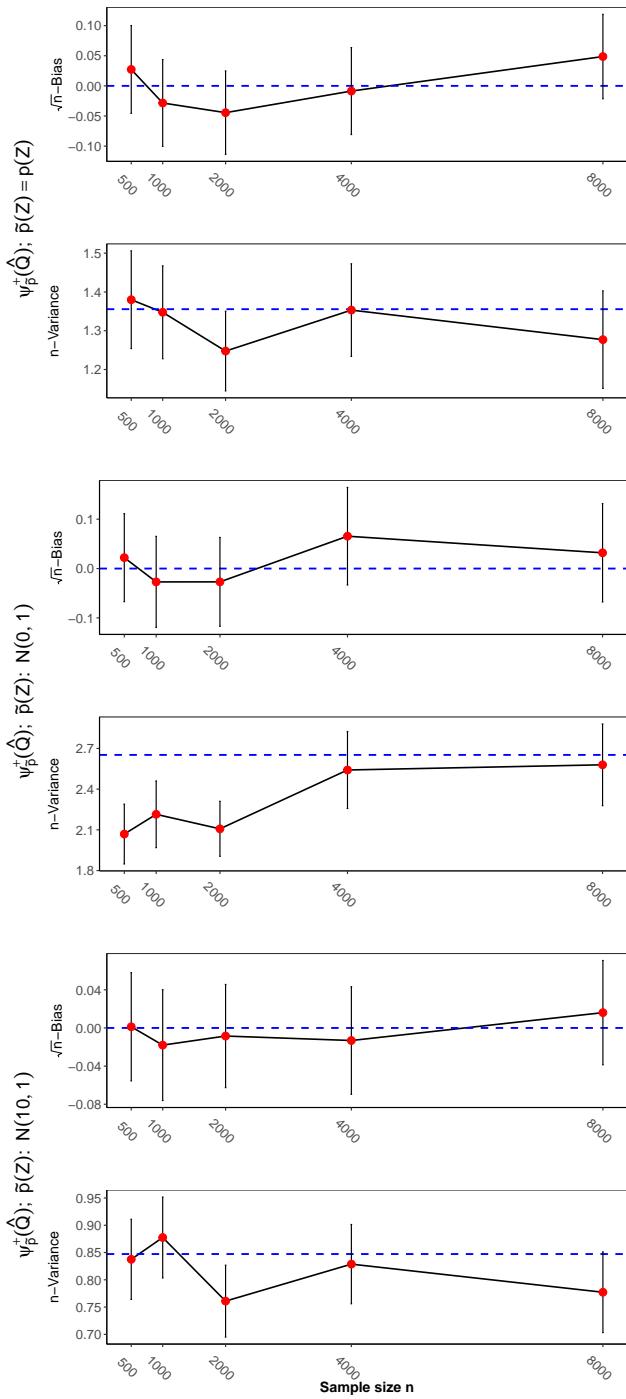


Figure 15: Simulation results demonstrating efficiency gains in ATE estimation when utilizing the Verma constraint under continuous Z .

H Details on real data application

H.1 Effect of mobile stroke unit care on functional outcomes

In [Piccininni et al. \[2023\]](#), only single mediator variable M_2 was adopted, and M_2 was categorized into three categories for easier estimation. Categorization was achieved using M_2 's first quantile and median value as (1) ≤ 48 minutes (1st quantile), (2) $48 - 75$ minutes (between 1st quantile and median), and (3) > 75 minutes (median) or no thrombolysis received. To compare with their result, we conducted the analysis under various scenarios by handling the outcome and mediator variables in various ways. First, we treated M_2 the categorized M_2 as continuous and Y as continuous. Then, we binarized them using different cut-off points. For the outcome Y , we applied three cut-off points: 2 (slight disability), 3 (moderate disability), and 4 (moderately severe disability). The binarization was as follows: (1) slight disability or less ($Y = 0$) vs. worse than slight disability ($Y = 1$), (2) moderate disability or less ($Y = 0$) vs. worse than moderate disability, and (3) moderately severe disability or less ($Y = 0$) vs. worse than moderately severe disability ($Y = 1$). For the mediator M_2 , we used two cut-offs: (1) ≤ 48 minutes ($M = 0$) vs. > 48 minutes ($M = 1$), and (2) ≤ 75 minutes ($M = 0$) vs. > 75 minutes or no thrombolysis received ($M = 1$). This resulted in six binary outcome-mediator scenarios.

We employed both the one-step estimator $\psi_1(\hat{Q})$ and TMLE $\psi_1(\hat{Q}^*)$ for ATE estimation when M_2 was binarized, and employed both $\psi_{2b}(\hat{Q})$ and TMLE $\psi_{2b}(\hat{Q}^*)$ when M_2 was treated as continuous. Super learner with five-fold cross-fitting was adopted to account for potential complex relationships among variables, such as interactions and nonlinear relationships. The super learner's candidate algorithms included intercept-only regression, generalized linear models, multivariate adaptive regression splines, and random forests. Missing data was handled with 10-fold multiple imputations.

Our analysis suggests that adopting MSU care is beneficial for improving patients' 3-month functional outcomes, aligned with the conclusions of [Piccininni et al. \[2023\]](#). This conclusion holds across different approaches to handling M and Y , although the statistical significance of the results varies. The beneficial effect of MSU care appears more pronounced in reducing mild disabilities, as indicated by larger effect sizes observed under lower cutoff values of Y . TMLE and one-step estimators yielded consistent and comparable results across all analyses.

Table 15: One-step and TMLE estimates of the average causal effect of additional mobile stroke unit (MSU) care on modified Rankin scale (mRS) score

<i>M</i> and <i>Y</i> type	<i>M</i> cutoff	<i>Y</i> cutoff	One-step estimator	TMLE
Continuous	-	-	-0.031, 95%CI (-0.4, 0.339)	-0.048, 95%CI (-0.465, 0.368)
Binary	48 mins	2	-0.046, 95%CI (-0.084, -0.009)	-0.048, 95%CI (-0.084, -0.012)
Binary	48 mins	3	-0.024, 95%CI (-0.062, 0.014)	-0.028, 95%CI (-0.063, 0.007)
Binary	48 mins	4	0, 95%CI (-0.035, 0.036)	-0.004, 95%CI (-0.036, 0.027)
Binary	75 mins	2	-0.031, 95%CI (-0.053, -0.008)	-0.035, 95%CI (-0.058, -0.012)
Binary	75 mins	3	-0.033, 95%CI (-0.058, -0.008)	-0.036, 95%CI (-0.061, -0.01)
Binary	75 mins	4	-0.015, 95%CI (-0.036, 0.005)	-0.017, 95%CI (-0.037, 0.004)

* Adopt one-step estimator $\psi_1(\hat{Q})$ and TMLE $\psi_1(\hat{Q}^*)$ under binary M , and adopt one-step estimator $\psi_{2b}(\hat{Q})$ and TMLE $\psi_{2b}(\hat{Q}^*)$ under continuous M .

H.2 Effect of academic performance on future annual income

Utilizing our front-door estimation framework, we investigated how early academic achievements influence future annual income. The data for this analysis was sourced from the Life Course Study, which spans from 1971 to 2002 and are publicly available through the Finnish Social Science Data Archive [Jorma, 2018]. These data originate from a longitudinal study of 634 individuals born between 1964 and 1968 in Jyväskylä, Finland. The study aimed to understand how abilities, social background, and educational achievements shape an individual’s life path. The data collection occurred in four phases. The first phase in the 1970s gathered initial information such as age, gender, family socioeconomic status, and results from the Illinois Test of Psycholinguistic Abilities (ITPA), assessing verbal intelligence in Finnish children aged 3-9. The second phase in the 1980s focused on academic achievements and performance. In 1991, the third phase collected data on occupational progress and higher education choices of the participants. Finally, the 2002 phase, as the subjects neared middle age, involved collecting information on their income, educational levels, and occupational status.

We were interested in estimating the causal effect of early academic performance (A) on an individual’s annual income (Y). We used a binary measure of academic performance based on whether an individual’s sixth-grade all-subject grade averages were above or below the median for the population. Our hypothesis is that early academic performance influences annual income by shaping educational and career paths, quantifiable through eight mediators ($M_1 - M_8$), detailed in Table 16. We also controlled for family socio-economic status, intelligence (measured by ITPA score), age, and gender ($X_1 - X_4$).

Table 16: Variable descriptions used in real data analysis (from the Finnish Social Science Data Archive.) Summary statistics contain information about mean and standard deviation for continuous variables and category frequency for categorical variables.

Variable	Definition; Summary statistic	Year
X_1	Socio-economic status as the total family taxable income in years 1983-84; 21619.54 (9806.7)	1983-84
X_2	ITPA score; 35.87 (5.97)	1971-72
X_3	Gender; male (49.68%), female (50.32%)	1971-91
X_4	Age; 25.17 (1.2)	1991
A	6th-grade all-subject grade averages compared to median; above (44.95%), below (55.05%)	1984
M_1	Undergraduate degree; yes (24.13%), no (75.87%)	1991
M_2	Highest educational field (categorised in accordance with Statistics Finland's Classification of Education 1988); science (90.06%), art (9.94%)	1991
M_3	Age at the start of the highest attained educational qualification; 19.33 (2.53)	1991
M_4	Length of formal education in months after comprehensive/upper secondary school (including education in progress; 28.55 (14.62))	1991
M_5	Number of different fields of education (including education in progress); 1.14 (0.5)	1991
M_6	Educational qualification required for current job; no (22.56%), somewhat (19.87%), yes (57.57%)	1991
M_7	Total length of the spells of unemployment greater than one year; no (84.07%), yes (15.93%)	1991
M_8	Age when started working; 21.34(2.4)	1991
Y	Respondent's earned income in euros in year 2000; 20541.93 (14462.12)	2002

Given the dimension of the mediators and due to the fact that the mediators include binary, categorical, and continuous-valued variables, we elected to use our proposed estimators that avoid mediator density estimation. Due to the potential for interactions and non-linear relationships, we wished to estimate nuisance parameters flexibly, and thus adopted a super learner approach combined with 5 folds cross-fitting. The candidate estimators included in the super learner include intercept-only regression, generalized linear models, multivariate adaptive regression splines, random forests, and XGBoost. For simplicity, we managed missing data in the variables mentioned by employing single imputation.

Our analysis underscores the role of strong academic foundations in shaping future income, likely mediated through higher educational attainment and more advantageous career paths. However, the interpretation of these estimates depends on the validity of the no direct effect assumption—namely, that the effect of academic performance on income operates entirely through the eight measured mediators (M_1 – M_8). Due to the lack of a valid anchor variable in this application, we cannot empirically test the front-door assumptions required for identifying the ATE. As such, we present two interpretations based on whether the no direct effect assumption is

believed to hold.

If the assumption holds, our TMLE estimator $\psi_{2b}(\hat{Q}^*)$ indicates that individuals with above-median academic performance in early stages earn, on average, €3239.18 more in future annual income (95% CI: €725.35, €5753.00) than their below-median counterparts. The one-step estimator $\psi_{2b}^+(\hat{Q}^*)$ provides a similar estimate of €3378.29 (95% CI: €857.74, €5898.84).

If the full mediation assumption (i.e., no direct effect of A on Y) is violated—such as when academic performance influences income through unmeasured pathways—then the ATE is not identifiable and the reported effects may be biased. In such cases, one can instead consider the PIIE estimand, which captures the effect of shifting the observed mediators under an intervention (see Section 2). The TMLE estimate suggests that shifting everyone’s educational and career paths to the values they would have taken under above-median academic performance would increase the average income by €1380.37 (95% CI: €-360.72, €3121.45), relative to the observed average income. Conversely, shifting everyone to the mediator values corresponding to below-median academic performance would decrease average income by €1858.81 (95% CI: €596.41, €3121.22), compared to the observed average income. The one-step estimator yields similar results, with an estimated increase of €1529.63 (95% CI: €-257.07, €3316.33) under above-median academic performance and a decrease of €1848.66 (95% CI: €638.75, €3058.57) under below-median performance.