Conditional Generative Adversarial Network for Individualized Causal Mediation Analysis with Survival Outcome

Cheng Huan, Xinyuan Song, and Hongwei Yuan

The Chinese University of Hong Kong and University of Macau

Supplementary Material

S1 Explanation of General Assumptions

Remark 1. Explanation of general assumptions:

(Ia) This assumption means the treatment assigned to one unit does not directly affect the outcomes of other units. In the context of survival outcomes, it implies the presence or absence of treatment for one individual does not impact the survival time or outcome of other individuals.

(Ib) This assumption means the treatment assigned to a unit has a consistent causal effect on that unit's outcome. In the context of survival outcomes, it implies the treatment has a constant impact on each individual's survival time or hazard rate, regardless of the treatment assignments of other individuals. By assuming both the stability of units and consistent

treatment effects, SUTVA allows us to leverage the presence of multiple units in estimating causal effects with survival outcomes. This feature enables us to draw valid causal inferences and evaluate the treatment effect on survival outcomes by comparing the outcomes of treated units with those of control units, accounting for potential confounding factors.

(II) This assumption plays a crucial role in the identification of causal effects in this study. Specifically, it ensures the validity of the identification formula (equation (2.4)) introduced in Section 2.3. The positivity of $P(M(t) = m|T = t, \mathbf{X} = \mathbf{x}) > 0$ is required to ensure that the integral is well-defined and that the conditional expectation $\mathbb{E}[Y(\mathbf{x}, t', M_t(\mathbf{x}))]$ can be estimated. Additionally, this assumption ensures that every subject has a positive probability of being assigned to each treatment arm, which is critical for making valid causal inferences. In the context of this study, as described in Section 7, the treatment variable T is binary, indicating the presence of APOE-ϵ4 alleles (1 = presence). The ADNI dataset includes 718 subjects, with 367 in the $T = 1$ group and 351 in the $T = 0$ group. This relatively balanced distribution ensures that $P(T = t|\mathbf{X} = \mathbf{x}) > 0$ is plausible across all levels of the covariates $\mathbf{X}$. The mediator M, defined as the difference in the proportion of ventricle volume in the whole brain between the 12th month and the baseline, is

a continuous variable that has been standardized prior to analysis. The standardization ensures that M has a smooth and continuous distribution, making $P(M(t) = m|T = t, \mathbf{X} = \mathbf{x}) > 0$ generally plausible across its range.

(III) The unconfoundedness assumptions guarantee the identification of ICEs with survival outcomes, as introduced in Section 2.3. These assumptions require that the treatment is independent of the potential outcomes and potential mediators, given the observed covariates $\mathbf{X}$, and the mediator is independent of the potential outcomes given the observed treatment T and pre-treatment covariates $\mathbf{X}$. By assuming treatment independence and mediator independence, the unconfoundedness assumption aims to ensure that any observed associations between treatment, mediator, and survival outcomes can be causally attributed to the causal effect and not to unmeasured confounders or reverse causality.

(IV) The assumption of noninformative censoring states that the censoring time C and the event time Y are conditionally independent given the covariates $\mathbf{X}$, treatment T, and mediator M. It is crucial for making valid causal inferences in survival analysis, ensuring that the censoring process is not related to the unobserved outcomes, given the treatment, mediator, and the observed covariates. As a result, we can treat the censored observations

as missing at random (MAR).

S2 Some Details in Section 3

S2.1 Mediator Layer

Counterfactual block: it consists of a generator and a discriminator. We first introduce the generator, $\mathbf{G_M} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \mapsto \mathcal{M} \times \mathcal{M}$, which takes several inputs, including the covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$), the binary treatment variable t (where $T = t$), the factual mediator m (where $M = m$), and some noise $\mathbf{Z}$. The output is the complete mediator vector $\mathbf{G_M}(\mathbf{Z}, \mathbf{x}, t, m) = \big(G_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{x}, t, m), G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{x}, t, m)\big)$, where $\mathbf{G_M}(\mathbf{Z}, \mathbf{X}, T, M) = \big(G_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{X}, T, M), G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{X}, T, M)\big)$ represents the random variable generated by $\mathbf{G_M}$. The discriminator, $D_{\mathbf{M}} : \mathcal{X} \times \mathcal{M} \times \mathcal{M} \mapsto [0,1]$, takes $\mathbf{x}$, $(1-t)m + tG_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{x}, t, m)$, and $tm + (1-t)G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{x}, t, m)$ as inputs. It outputs a scalar value representing the probability that the last input $tm + (1-t)G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{x}, t, m)$ corresponds to the factual mediator rather than the counterfactual mediator. This setup allows us to generate the complete mediator vector by incorporating the covariates, treatment, factual mediator, and noise into the generator, and then leveraging the discriminator to distinguish between the factual and counterfactual mediators. The loss function associated with this setup is

$$\mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}) := \mathbb{E}_{(\mathbf{X},T,M)\sim P_{\mathbf{X},T,M}} \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}} \Big\{ T \log D_{\mathbf{M}}\big(\mathbf{X}, (1-T)M + TG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M)\big) + (1-T)\log\big[1 - D_{\mathbf{M}}\big(\mathbf{X}, (1-T)M + TG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M)\big)\big]\Big\},$$

and the corresponding minimax optimization problem can be formulated as $\min_{\mathbf{G}_{\mathbf{M}}} \max_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})$. At the population level, the target conditional generator and discriminator, $\mathbf{G}_{\mathbf{M}}^*$ and $D_{\mathbf{M}}^*$, are defined as the solutions to the optimization problem:

$$(\mathbf{G}_{\mathbf{M}}^*, D_{\mathbf{M}}^*) = \text{argmin}_{\mathbf{G}_{\mathbf{M}}} \text{argmax}_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}). \tag{S2.1}$$

Similar to $\mathbf{G}_{\mathbf{M}} = (G_{\mathbf{M}}^{(0)}, G_{\mathbf{M}}^{(1)})$, we denote $\mathbf{G}_{\mathbf{M}}^* = (G_{\mathbf{M}}^{*,(0)}, G_{\mathbf{M}}^{*,(1)})$.

Empirical Loss Function of Counterfactual Block: for the dataset $\{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i\}_{i=1}^n$, independently and identically distributed according to $P_{\mathbf{X},T,M}$, and $\{\mathbf{Z} = \mathbf{z}_i\}_{i=1}^n$ independently generated from $P_{\mathbf{Z}}$, we define the sample set $S_n^M := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \mathbf{Z} = \mathbf{z}_i\}_{i=1}^n$. This sample set is used to train the estimated conditional generator $\widehat{\mathbf{G}}_{\mathbf{M}}$. We consider the empirical version of $\mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})$:

$$\widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}) = \frac{1}{n}\sum_{i=1}^n \Big\{ t_i \log D_{\mathbf{M}}\Big(\mathbf{x}_i, (1-t_i)m_i + t_i G_{\mathbf{M}}^{(0)}(\mathbf{z}_i,\mathbf{x}_i,t_i,m_i), t_i m_i + (1-t_i)G_{\mathbf{M}}^{(1)}(\mathbf{z}_i,\mathbf{x}_i,t_i,m_i)\Big) + (1-t_i)\log\Big[1 - D_{\mathbf{M}}\Big(\mathbf{x}_i, (1-t_i)m_i + t_i G_{\mathbf{M}}^{(0)}(\mathbf{z}_i,\mathbf{x}_i,t_i,m_i), t_i m_i + (1-t_i)G_{\mathbf{M}}^{(1)}(\mathbf{z}_i,\mathbf{x}_i,t_i,m_i)\Big)\Big]\Big\}.$$

We also introduce a supervised loss to ensure that $G_{\mathbf{M}}^{(t)}(\mathbf{z}, \mathbf{x}, t, m) = m$: $\widetilde{\mathcal{L}}_1(\mathbf{G_M}) := \frac{1}{n}\sum_{i=1}^{n}\left|G_{\mathbf{M}}^{(t_i)}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i) - m_i\right|^2$. Now, we can define the empirical objective function as follows: for a supervised parameter $\alpha_1 \geq 0$,

$$\widehat{\mathcal{L}}_{\mathbf{M}}(\mathbf{G_M}, D_{\mathbf{M}}) := \widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G_M}, D_{\mathbf{M}}) + \alpha_1\widetilde{\mathcal{L}}_1(\mathbf{G_M}). \tag{S2.2}$$

We use two feedforward neural networks (FNN) to estimate $\mathbf{G_M}$, denoted as $\widehat{\mathbf{G}}_{\mathbf{M}}$, based on (S2.2). See details in Section S2.3.

Inferential Block: we extend the classical CGAN framework in this block to generate the complete mediator vector solely based on the given covariates $\mathbf{x}$, without relying on factual mediator and treatment. After training the above counterfactual mediator block, we obtain the complete mediator vector $\big((1-t)m + tG_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{x}, t, m), tm + (1-t)G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{x}, t, m)\big)$. Then, we transfer this complete mediator vector and the given covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$) to the inferential mediator block for inference. The generator, denoted as $\mathbf{I_M} : \mathbb{R}^{d_z} \times \mathcal{X} \mapsto \mathcal{M} \times \mathcal{M}$, takes the covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$) and some noise $\widehat{\mathbf{Z}}$ as inputs. It produces the complete mediator vector $\mathbf{I_M}(\widehat{\mathbf{Z}}, \mathbf{x}) = \big(I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}}, \mathbf{x}), I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}}, \mathbf{x})\big)$, where $\mathbf{I_M}(\widehat{\mathbf{Z}}, \mathbf{X}) = \big(\mathbf{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}}, \mathbf{X}), \mathbf{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}}, \mathbf{X})\big)$ represents the random variable generated by $\mathbf{I_M}$. The discriminator, $D_{\mathbf{I_M}}$, in this case, takes either $\big(\mathbf{x}, (1-t)m + tG_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{x}, t, m), tm + (1-t)G_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{x}, t, m)\big)$ or $\big(\mathbf{x}, \mathbf{I_M}(\widehat{\mathbf{z}}, \mathbf{x})\big)$ as inputs. By utilizing this architecture, we can generate the complete mediator vector by integrating the covariates and noise into the generator. The discriminator helps distinguish between the counterfactual and inferred complete mediator vectors.

We employ the classical CGAN loss:

$$\mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}) = \mathbb{E}_{\mathbf{q}\sim P_{\mathbf{Q}}}[\log D_{\mathbf{I_M}}(\mathbf{q})] + \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\mathbb{E}_{\widehat{\mathbf{z}}\sim P_{\widehat{\mathbf{Z}}}} \log\Big[1 - D_{\mathbf{I_M}}\big(\mathbf{x}, I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}, \mathbf{x}), I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}, \mathbf{x})\big)\Big],$$

where $P_{\mathbf{Q}}$ is the the joint distribution of $(\mathbf{X}, G_{\mathbf{M}}^{*,(0)}(\mathbf{Z}, \mathbf{X}, T = 1, M_1(\mathbf{X})), M_1(\mathbf{X}))$. Subsequently, we aim to solve $\min_{\mathbf{I_M}} \max_{D_{\mathbf{I_M}}} \mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}})$. Define

$$(\mathbf{I}_{\mathbf{M}}^*, D_{\mathbf{I_M}}^*) := \text{argmin}_{\mathbf{I_M}}\text{argmax}_{D_{\mathbf{I_M}}} \mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}) \text{ and } \mathbb{L}_{\mathbf{IM}}(\mathbf{I_M}) := \sup_{D_{\mathbf{I_M}}} \mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}),$$

and denote $\mathbf{I}_{\mathbf{M}}^* = (I_{\mathbf{M}}^{*,(0)}, I_{\mathbf{M}}^{*,(1)})$. Based on the Lemmas regarding distribution matching presented in Section S2.2, we can show that $I_{\mathbf{M}}^{*,(0)}(\widehat{\mathbf{Z}}, \mathbf{X}) \sim M_0(\mathbf{X}) \sim P_{M|\mathbf{X},T=0}$ and $I_{\mathbf{M}}^{*,(1)}(\widehat{\mathbf{Z}}, \mathbf{X}) \sim M_1(\mathbf{X}) \sim P_{M|\mathbf{X},T=1}$.

Empirical Loss Function of Inferential Block: given the sample set S_n^M and $\{\widehat{\mathbf{Z}} = \widehat{\mathbf{z}}_i\}_{i=1}^n$ independently generated from $P_{\widehat{\mathbf{Z}}}$, after obtaining $\widehat{\mathbf{G}}_{\mathbf{M}}$ in the previous section, we define another sample set $S_n^{IM} := \{(\mathbf{x}_i, t_i, \overline{m}_i^{(0)}, \overline{m}_i^{(1)}, \widehat{\mathbf{z}}_i)\}_{i=1}^n$, where $(\overline{m}_i^{(0)}, \overline{m}_i^{(1)}) = t_i\big(\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T = 1, m_i), m_i\big) + (1 - t_i)\big(m_i, \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T = 0, m_i)\big)$, to train the estimated conditional generator $\widehat{\mathbf{I}}_{\mathbf{M}}$ in the inferential block. Consider the following empirical version of $\mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}})$:

$$\widetilde{\mathcal{L}}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}; \widehat{\mathbf{G}}_{\mathbf{M}}) = \frac{1}{n}\sum_{i=1}^n \Big\{ \log D_{\mathbf{I_M}}\big(\mathbf{x}_i, \overline{m}_i^{(0)}, \overline{m}_i^{(1)}\big) + \log\Big[1 - D_{\mathbf{I_M}}\big(\mathbf{x}_i, I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i), I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i)\big)\Big]\Big\}.$$

To optimize the performance with respect to equation $\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\Big[\big|\mathbb{E}\big[M_1(\mathbf{x}) - M_0(\mathbf{x})\big] - \mathbb{E}\big[I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}}, \mathbf{x}) - I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}}, \mathbf{x})\big]\big|^2\Big]$, we introduce a supervised loss: $\widetilde{\mathcal{L}}_2(\mathbf{I_M}; \widehat{\mathbf{G}}_{\mathbf{M}}) = \frac{1}{n}\sum_{i=1}^n \big|\big(\overline{m}_i^{(0)} - \overline{m}_i^{(1)}\big) - \big(I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i) - I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i)\big)\big|^2$. Meanwhile, we introduce another supervised loss to ensure $I_{\mathbf{M}}^{(t)}(\mathbf{z}, \mathbf{x}) = m$: $\widetilde{\mathcal{L}}_3(\mathbf{I_M}) = \frac{1}{n}\sum_{i=1}^n \big|I_{\mathbf{M}}^{(t_i)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i) - m_i\big|^2$. For supervised parameters α_2, $\alpha_3 \geq 0$, we define an empirical objec-

tive function as follows:

$$\widehat{\mathcal{L}}_{\mathbf{IM}}(\mathbf{I}_{\mathbf{M}}, D_{\mathbf{I}_{\mathbf{M}}}; \widehat{\mathbf{G}}_{\mathbf{M}}) := \widetilde{\mathcal{L}}_{\mathbf{IM}}(\mathbf{I}_{\mathbf{M}}, D_{\mathbf{I}_{\mathbf{M}}}; \widehat{\mathbf{G}}_{\mathbf{M}}) + \alpha_2 \widetilde{\mathcal{L}}_2(\mathbf{I}_{\mathbf{M}}; \widehat{\mathbf{G}}_{\mathbf{M}}) + \alpha_3 \widetilde{\mathcal{L}}_3(\mathbf{I}_{\mathbf{M}}). \tag{S2.3}$$

Again, we use FNN to estimate $\mathbf{I}_{\mathbf{M}}$, denoted as $\widehat{\mathbf{I}}_{\mathbf{M}}$, based on the empirical objective function (S2.3). See details in Section S2.4.

S2.2 Distribution Matching

First, we consider the loss function

$$\begin{aligned}
&\mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}) := \\
&\mathbb{E}_{(\mathbf{X},T,M)\sim P_{\mathbf{X},T,M}} \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}} \Big\{ T \log D_{\mathbf{M}}\Big(\mathbf{X}, (1-T)M + TG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M)\Big) \\
&\qquad + (1-T)\log[1 - D_{\mathbf{M}}\Big(\mathbf{X}, (1-T)M + TG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M)\Big)]\Big\},
\end{aligned}$$

and

$$(\mathbf{G}_{\mathbf{M}}^*, D_{\mathbf{M}}^*) = \mathrm{argmin}_{\mathbf{G}_{\mathbf{M}}} \mathrm{argmax}_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}). \tag{S2.4}$$

Now, we provide the form of the target discriminator $D_{\mathbf{M}}^*$. For any measurable function $\mathbf{G}_{\mathbf{M}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \mapsto \mathcal{M} \times \mathcal{M}$, we define $\mathbb{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}) := \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})$. Then,

$$\begin{aligned}
\mathbb{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}) &= \sup_{D_{\mathbf{M}}} \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}} \Big\{ P(T=1|\mathbf{X}) \cdot \mathbb{E}_{M\sim P_{M|\mathbf{X},T=1}} \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}} \big\{ \log D_{\mathbf{M}}(\mathbf{X}, G_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},1,M), M)\big\} \\
&\qquad + P(T=0|\mathbf{X}) \cdot \mathbb{E}_{M\sim P_{M|\mathbf{X},T=0}} \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}} \big\{ \log[1 - D_{\mathbf{M}}(\mathbf{X}, M, G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},0,M))]\big\} \Big\} \\
&= \sup_{D_{\mathbf{M}}} \int_{\mathcal{X}} \int_{\mathcal{M}} \int_{\mathbb{R}^{d_z}} \Big(\log D_{\mathbf{M}}(\mathbf{x}, G_{\mathbf{M}}^{(0)}(\mathbf{z},\mathbf{x},1,m), m) \cdot p_{M|\mathbf{X},T}(M=m|\mathbf{X}=\mathbf{x}, T=1) \cdot p_{T|\mathbf{X}}(T=1|\mathbf{X}=\mathbf{x}) \\
&\qquad + \log[1 - D_{\mathbf{M}}(\mathbf{x}, m, G_{\mathbf{M}}^{(1)}(\mathbf{z},\mathbf{x},0,m))] \cdot p_{M|\mathbf{X},T}(M=m|\mathbf{X}=\mathbf{x}, T=0) \cdot p_{T|\mathbf{X}}(T=0|\mathbf{X}=\mathbf{x}) \Big) \\
&\qquad \cdot p_{\mathbf{Z}}(\mathbf{Z}=\mathbf{z})\, d\mathbf{z}\, dm\, p_{\mathbf{X}}(\mathbf{X}=\mathbf{x}) d\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
=&\sup_{D_{\mathbf{M}}} \int_{\mathcal{X}} \int_{\mathcal{M}} \int_{\mathbb{R}^{d_z}} \Big(\log D_{\mathbf{M}}(\mathbf{x}, G_{\mathbf{M}}^{(0)}(\mathbf{z}, \mathbf{x}, 1, m), m) \cdot p_{\mathbf{Z},\mathbf{X},T,M}(\mathbf{Z}=\mathbf{z}, \mathbf{X}=\mathbf{x}, T=1, M=m) \\
&\qquad + \log[1 - D_{\mathbf{M}}(\mathbf{x}, m, G_{\mathbf{M}}^{(1)}(\mathbf{z}, \mathbf{x}, 0, m))] \cdot p_{\mathbf{Z},\mathbf{X},T,M}(\mathbf{Z}=\mathbf{z}, \mathbf{X}=\mathbf{x}, T=0, M=m) \Big) \, d\mathbf{z} \, dm \, d\mathbf{x} \\
=&\sup_{D_{\mathbf{M}}} \Bigg(\int_{G_{\mathbf{M},1}(\mathbb{R}^{d_z} \times \mathcal{X} \times \{1\} \times \mathcal{M})} \log D_{\mathbf{M}}(\mathbf{q}) \cdot dP_{\mathbf{Z},\mathbf{X},T,M}\Big((\mathbf{Z}, \mathbf{X}, T, M) \in G_{\mathbf{M},1}^{-1}(\mathbf{q})\Big) \\
&\quad + \int_{G_{\mathbf{M},0}(\mathbb{R}^{d_z} \times \mathcal{X} \times \{0\} \times \mathcal{M})} \log[1 - D_{\mathbf{M}}(\mathbf{w})] \cdot dP_{\mathbf{Z},\mathbf{X},T,M}\Big((\mathbf{Z}, \mathbf{X}, T, M) \in G_{\mathbf{M},0}^{-1}(\mathbf{w})\Big) \Bigg) \\
=&\sup_{D_{\mathbf{M}}} \Big(\mathbb{E}_{\mathbf{q} \sim P_{\mathbf{Q}}} \log D_{\mathbf{M}}(\mathbf{q}) + \mathbb{E}_{\mathbf{w} \sim P_{\mathbf{W}}} \log[1 - D_{\mathbf{M}}(\mathbf{w})] \Big) \\
=&\mathbb{D}_{JS}(p_{\mathbf{Q}}, p_{\mathbf{W}}) - \log 4 = \mathbb{D}_{KL}(p_{\mathbf{Q}} \| (p_{\mathbf{Q}} + p_{\mathbf{W}})/2) + \mathbb{D}_{KL}(p_{\mathbf{W}} \| (p_{\mathbf{Q}} + p_{\mathbf{W}})/2) - \log 4, \qquad \text{(S2.5)}
\end{aligned}
$$

where $p_{\mathbf{Q}}$ and $p_{\mathbf{W}}$ are densities of $P_{\mathbf{Q}}(\mathbf{Q}=\mathbf{q}) = P_{\mathbf{Z},\mathbf{X},T,M}\Big((\mathbf{Z}, \mathbf{X}, T, M) \in G_{\mathbf{M},1}^{-1}(\mathbf{q})\Big)$ for any $\mathbf{q} \in G_{\mathbf{M},1}(\mathbb{R}^{d_z} \times \mathcal{X} \times \{1\} \times \mathcal{M})$ and $P_{\mathbf{W}}(\mathbf{W}=\mathbf{w}) = P_{\mathbf{Z},\mathbf{X},T,M}\Big((\mathbf{Z}, \mathbf{X}, T, M) \in G_{\mathbf{M},0}^{-1}(\mathbf{w})\Big)$ for any $\mathbf{w} \in G_{\mathbf{M},0}(\mathbb{R}^{d_z} \times \mathcal{X} \times \{0\} \times \mathcal{M})$, respectively; $G_{\mathbf{M},1} : (\mathbf{z}, \mathbf{x}, T, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times \{1\} \times \mathcal{M} \mapsto (\mathbf{x}, G_{\mathbf{M}}^{(0)}(\mathbf{z}, \mathbf{x}, 1, m), m) \in \mathcal{X} \times \mathcal{M} \times \mathcal{M}$, and $G_{\mathbf{M},0} : (\mathbf{z}, \mathbf{x}, T, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times \{0\} \times \mathcal{M} \mapsto (\mathbf{x}, m, G_{\mathbf{M}}^{(1)}(\mathbf{z}, \mathbf{x}, 0, m)) \in \mathcal{X} \times \mathcal{M} \times \mathcal{M}$; $G_{\mathbf{M},i}^{-1}$ is the inverse mapping of $G_{\mathbf{M},i}$ for $i \in \{0, 1\}$; $\mathbb{D}_{JS}(A, B)$ denotes the Jensen-Shannon (JS) divergence between two distributions A and B, and $\mathbb{D}_{KL}(A\|B)$ denotes the Kullback-Leibler (KL) divergence between A and B.

Using the properties of the f-divergence including JS divergence as a special case (Zhou et al., 2022), we obtain that the optimal discriminator is

$$
D_{\mathbf{M}}^{*} = \frac{p_{\mathbf{Q}}}{p_{\mathbf{Q}} + p_{\mathbf{W}}} = \frac{p_{\mathbf{X},G_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}}{p_{\mathbf{X},G_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} + p_{\mathbf{X},M_0(\mathbf{X}),G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}}. \qquad \text{(S2.6)}
$$

According to Statement 1 of Theorem S.2 in Arjovsky et al. (2017), we can deduce that $\mathbb{D}_{JS}(p_{\mathbf{Q}}, p_{\mathbf{W}}) = 0$ if and only if the total variation distance of

probability measures is zero, i.e., $\|P_{\mathbf{Q}} - P_{\mathbf{W}}\|_{TV} := \sup_{A\in\Sigma_{\mathcal{X}\times\mathcal{M}\times\mathcal{M}}} |P_{\mathbf{Q}}(A) - P_{\mathbf{W}}(A)| = \frac{1}{2}\|p_{\mathbf{Q}} - p_{\mathbf{W}}\|_{L^1} := \frac{1}{2}\int_{\mathcal{X}\times\mathcal{M}\times\mathcal{M}} |p_{\mathbf{Q}}(\mathbf{q}) - p_{\mathbf{W}}(\mathbf{q})|d\mathbf{q} = 0$. Here, $\Sigma_{\mathcal{X}\times\mathcal{M}\times\mathcal{M}}$ represents the set of all measurable subsets, $P_{\mathbf{Q}}(A) = \int_{\mathbf{q}\in A} p_{\mathbf{Q}}(\mathbf{q})d\mathbf{q}$, $P_{\mathbf{W}}(A) = \int_{\mathbf{q}\in A} p_{\mathbf{W}}(\mathbf{q})d\mathbf{q}$, and the second equality can be verified using Proposition 4.2 in Levin and Peres (2017). As a consequence, we obtain the following lemma on distribution matching:

Lemma 1. *A function* $\mathbf{G}^*_{\mathbf{M}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \mapsto \mathcal{M} \times \mathcal{M}$ *is a minimizer of* $\mathbb{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}})$*, that is,* $\mathbf{G}^*_{\mathbf{M}} \in argmin_{\mathbf{G}_{\mathbf{M}}}\mathbb{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}})$*, if and only if*

$$\left\| p_{\mathbf{X},G^{*,(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},M_0(\mathbf{X}),G^{*,(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} \right\|_{L^1} = 0;$$

that is, $\left(\mathbf{X}, G^{*,(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})), M_1(\mathbf{X})\right) \sim \left(\mathbf{X}, M_0(\mathbf{X}), G^{*,(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))\right)$.

Denote the joint distribution in the above lemma by $P_{\mathbf{Q}}$.

Drawing an analogy to the counterfactual block in the mediator layer, we can derive the expression for the optimal discriminator $D^*_{\mathbf{Y}}$ in the outcome layer. For any measurable function $\mathbf{G}_{\mathbf{Y}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \times \mathcal{Y}$, the optimal discriminator is given by:

$$D^*_{\mathbf{Y}} = \frac{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)}}{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)} + p_{\mathbf{X},M,Y_0(\mathbf{X},M),G^{(1)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))}}. \quad \text{(S2.7)}$$

We have the following lemma regarding distribution matching.

Lemma 2. *A function* $\mathbf{G}^*_{\mathbf{Y}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \times \mathcal{Y} \mapsto \mathcal{Y} \times \mathcal{Y}$ *is a minimizer of* $\mathbb{L}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}) := \sup_{D_{\mathbf{Y}}} \mathcal{L}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}, D_{\mathbf{Y}})$*, that is,* $\mathbf{G}^*_{\mathbf{Y}} \in argmin_{\mathbf{G}_{\mathbf{Y}}}\mathbb{L}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}})$*, if*

and only if

$$\left\| p_{\mathbf{X},M,G_{\mathbf{Y}}^{*,(0)}(\widetilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)} - p_{\mathbf{X},M,Y_0(\mathbf{X},M),G_{\mathbf{Y}}^{*,(1)}(\widetilde{\mathbf{Z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))} \right\|_{L^1} = 0,$$

or

$$\Big(\mathbf{X}, M, G_{\mathbf{Y}}^{*,(0)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T=1, M, Y_1(\mathbf{X}, M)), Y_1(\mathbf{X}, M)\Big) \sim \Big(\mathbf{X}, M, Y_0(\mathbf{X}, M), G_{\mathbf{Y}}^{*,(1)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T=0, M, Y_0(\mathbf{X}, M))\Big).$$

Denote the joint distribution in the above lemma as $P_{\widetilde{\mathbf{Q}}}$.

Next, consider the classical CGAN loss:

$$\mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}) = \mathbb{E}_{\mathbf{q}\sim P_{\mathbf{Q}}}[\log D_{\mathbf{I_M}}(\mathbf{q})] + \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\mathbb{E}_{\widehat{\mathbf{z}}\sim P_{\widehat{\mathbf{Z}}}} \log\Big[1 - D_{\mathbf{I_M}}\big(\mathbf{x}, I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}, \mathbf{x}), I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}, \mathbf{x})\big)\Big],$$

and

$$(\mathbf{I}_{\mathbf{M}}^*, D_{\mathbf{I_M}}^*) := \text{argmin}_{\mathbf{I_M}}\text{argmax}_{D_{\mathbf{I_M}}}\mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}) \text{ and } \mathbb{L}_{\mathbf{IM}}(\mathbf{I_M}) := \sup_{D_{\mathbf{I_M}}} \mathcal{L}_{\mathbf{IM}}(\mathbf{I_M}, D_{\mathbf{I_M}}).$$

Based on the standard theory of CGAN (Goodfellow et al., 2014; Mirza and Osindero, 2014), it can be shown that the optimal discriminator is given by

$$D_{\mathbf{I_M}}^* = \frac{p_{\mathbf{Q}}}{p_{\mathbf{Q}} + p_{\mathbf{X},\mathbf{I_M}(\widehat{\mathbf{Z}},\mathbf{X})}}, \tag{S2.8}$$

and we can establish the following lemma regarding distribution matching.

Lemma 3. *A function* $\mathbf{I}_{\mathbf{M}}^* : \mathbb{R}^{d_z} \times \mathcal{X} \mapsto \mathcal{M} \times \mathcal{M}$ *is a minimizer of* $\mathbb{L}_{\mathbf{IM}}(\mathbf{I_M})$*, that is,* $\mathbf{I}_{\mathbf{M}}^* \in argmin_{\mathbf{I_M}}\mathbb{L}_{\mathbf{IM}}(\mathbf{I_M})$ *if and only if* $\Big\| p_{\mathbf{Q}} - p_{\mathbf{X},\mathbf{I_M}(\widehat{\mathbf{Z}},\mathbf{X})} \Big\|_{L^1} = 0$*; that is,* $\Big(\mathbf{X}, G_{\mathbf{M}}^{*,(0)}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})\Big) \sim \Big(\mathbf{X}, M_0(\mathbf{X}), G_{\mathbf{M}}^{*,(1)}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))\Big) \sim \Big(\mathbf{X}, I_{\mathbf{M}}^{*,(0)}(\widehat{\mathbf{Z}}, \mathbf{X}), I_{\mathbf{M}}^{*,(1)}(\widehat{\mathbf{Z}}, \mathbf{X})\Big)$.

By considering the marginal distributions, we can conclude that $I_{\mathbf{M}}^{*,(0)}(\widehat{\mathbf{Z}}, \mathbf{X}) \sim M_0(\mathbf{X}) \sim P_{M|\mathbf{X},T=0}$ and $I_{\mathbf{M}}^{*,(1)}(\widehat{\mathbf{Z}}, \mathbf{X}) \sim M_1(\mathbf{X}) \sim P_{M|\mathbf{X},T=1}$.

Similarly, by the standard theory of CGAN, we can determine the optimal discriminator of inferential block in the outcome layer as follows:

$$D_{\mathbf{I_Y}}^{*} = \frac{p_{\widetilde{\mathbf{Q}}}}{p_{\widetilde{\mathbf{Q}}} + p_{\mathbf{X},M,\mathbf{I_Y}(\widehat{\mathbf{Z}},\mathbf{X})}}, \tag{S2.9}$$

and obtain the following lemma on distribution matching.

Lemma 4. *A function* $\mathbf{I}_{\mathbf{Y}}^{*} : \mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M} \mapsto \mathcal{Y} \times \mathcal{Y}$ *is a minimizer of* $\mathbb{L}_{\mathbf{IY}}(\mathbf{I_Y})$*, that is,* $\mathbf{I}_{\mathbf{Y}}^{*} \in argmin_{\mathbf{I_Y}} \mathbb{L}_{\mathbf{IY}}(\mathbf{I_Y})$*, if and only if* $\|p_{\widetilde{\mathbf{Q}}} - p_{\mathbf{X},M,\mathbf{I_Y}(\overline{\mathbf{Z}},\mathbf{X},M)}\|_{L^1} = 0$*; that is,* $\left(\mathbf{X}, M, G_{\mathbf{Y}}^{*,(0)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T=1, M, Y_1(\mathbf{X}, M)), Y_1(\mathbf{X}, M)\right) \sim \left(\mathbf{X}, M, Y_0(\mathbf{X}, M), G_{\mathbf{Y}}^{*,(1)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T=0, M, Y_0(\mathbf{X}, M))\right) \sim \left(\mathbf{X}, M, I_{\mathbf{Y}}^{*,(0)}(\overline{\mathbf{Z}}, \mathbf{X}, M), I_{\mathbf{Y}}^{*,(1)}(\overline{\mathbf{Z}}, \mathbf{X}, M)\right)$.

By considering the marginal distributions, we deduce that $I_{\mathbf{Y}}^{*,(0)}(\overline{\mathbf{Z}}, \mathbf{X}, M) \sim Y_0(\mathbf{X}, M) \sim P_{Y|\mathbf{X},T=0,M}$ and $I_{\mathbf{Y}}^{*,(1)}(\overline{\mathbf{Z}}, \mathbf{X}, M) \sim Y_1(\mathbf{X}, M) \sim P_{Y|\mathbf{X},T=1,M}$.

S2.3 $\mathbf{G_M}$ estimation

We use two FNNs (Goodfellow et al., 2016) to estimate $\mathbf{G_M}$ based on the empirical objective function (3.11). Denote the conditional generator network as $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, and the conditional discriminator network as $D_{\mathbf{M}}^{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\phi}$. For any function $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^d$, denote $\|f\|_{L^\infty} = \sup_{\mathbf{x}\in\mathcal{X}} \|f(\mathbf{x})\|$, where $\|\cdot\|$ is the Euclidean norm.

- The generator network $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$: let $\mathcal{G} \equiv \mathcal{G}_{\mathcal{H},\mathcal{W},\mathcal{S},\mathcal{B}}$ be the set of ReLU

neural networks $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \mapsto \mathcal{M} \times \mathcal{M}$ parameterized by $\boldsymbol{\theta}$, depth $\mathcal{H}$, width $\mathcal{W}$, size $\mathcal{S}$, and $\left\|\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}\right\|_{L^\infty} \leq \mathcal{B}$. Here, the depth $\mathcal{H}$ refers to the number of hidden layers, so the network has $\mathcal{H}+1$ layers in total. A $(\mathcal{H}+1)$-vector $(w_0, w_1, \ldots, w_{\mathcal{H}})$ specifies the width of each layer, where w_0 is the dimension of the input data and $w_{\mathcal{H}}$ is the dimension of the output. The width $\mathcal{W} = \max\{w_1, \ldots, w_{\mathcal{H}}\}$ is the maximum width of the hidden layers. The size $\mathcal{S} = \sum_{i=0}^{\mathcal{H}} [w_i \times w_{i+1}]$ is the total number of parameters in the network.

- The discriminator network $D_{\mathbf{M}}^{\boldsymbol{\phi}}$: denote $\mathcal{D} \equiv \mathcal{D}_{\widetilde{\mathcal{H}},\widetilde{\mathcal{W}},\widetilde{\mathcal{S}},\widetilde{\mathcal{B}}}$ as the set of ReLU neural networks $D_{\mathbf{M}}^{\boldsymbol{\phi}} : \mathcal{X} \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$, parameterized by $\boldsymbol{\phi}$, depth $\widetilde{\mathcal{H}}$, width $\widetilde{\mathcal{W}}$, size $\widetilde{\mathcal{S}}$, and $\left\|D_{\mathbf{M}}^{\boldsymbol{\phi}}\right\|_{L^\infty} \leq \widetilde{\mathcal{B}}$.

Then, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated as follows: $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\phi}}) = \text{argmin}_{\boldsymbol{\theta}}\text{argmax}_{\boldsymbol{\phi}}\widehat{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}, D_{\mathbf{M}}^{\boldsymbol{\phi}})$, and the estimated conditional generator is $\widehat{\mathbf{G}}_{\mathbf{M}} = \mathbf{G}_{\mathbf{M}}^{\widehat{\boldsymbol{\theta}}}$, and the estimated discriminator is $\widehat{D}_{\mathbf{M}} = D_{\mathbf{M}}^{\widehat{\boldsymbol{\phi}}}$. Note that $\widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})$ depends on $G_{\mathbf{M}}^{(0)}(\cdot,\cdot,1,\cdot)$ and $G_{\mathbf{M}}^{(1)}(\cdot,\cdot,0,\cdot)$ but not on $G_{\mathbf{M}}^{(0)}(\cdot,\cdot,0,\cdot)$ or $G_{\mathbf{M}}^{(1)}(\cdot,\cdot,1,\cdot)$, while $\widetilde{\mathcal{L}}_1(\mathbf{G}_{\mathbf{M}})$ exhibits the opposite behavior, meaning that it is dependent on $G_{\mathbf{M}}^{(0)}(\cdot,\cdot,0,\cdot)$ and $G_{\mathbf{M}}^{(1)}(\cdot,\cdot,1,\cdot)$ but not on $G_{\mathbf{M}}^{(0)}(\cdot,\cdot,1,\cdot)$ or $G_{\mathbf{M}}^{(1)}(\cdot,\cdot,0,\cdot)$. Therefore, $\widehat{\mathbf{G}}_{\mathbf{M}}$ is also a minimizer of both $\widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}, \widehat{D}_{\mathbf{M}})$ and $\widetilde{\mathcal{L}}_1(\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}})$.

S2.4 $\mathbf{I_M}$ estimation

Denote the conditional generator network as $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}$ parameterized by $\boldsymbol{\psi}$, and the conditional discriminator network as $D_{\mathbf{I_M}}^{\boldsymbol{\omega}}$ parameterized by $\boldsymbol{\omega}$.

• The generator network $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}$: let $\mathcal{I} \equiv \mathcal{I}_{\overline{\mathcal{H}},\overline{\mathcal{W}},\overline{\mathcal{S}},\overline{\mathcal{B}}}$ be the set of ReLU neural networks $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}} : \mathbb{R}^{d_{\hat{z}}} \times \mathcal{X} \mapsto \mathcal{M} \times \mathcal{M}$ parameterized by $\boldsymbol{\psi}$, depth $\overline{\mathcal{H}}$, width $\overline{\mathcal{W}}$, size $\overline{\mathcal{S}}$, and $\left\|\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}\right\|_{L^\infty} \leq \overline{\mathcal{B}}$.

• The discriminator network $D_{\mathbf{I_M}}^{\boldsymbol{\omega}}$: denote $\mathcal{D}_I \equiv \mathcal{D}_{\overline{\mathcal{H}}_D,\overline{\mathcal{W}}_D,\overline{\mathcal{S}}_D,\overline{\mathcal{B}}_D}$ as the set of ReLU neural networks $D_{\mathbf{I_M}}^{\boldsymbol{\omega}} : \mathcal{X} \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$, parameterized by $\boldsymbol{\omega}$, depth $\overline{\mathcal{H}}_D$, width $\overline{\mathcal{W}}_D$, size $\overline{\mathcal{S}}_D$, and $\left\|D_{\mathbf{I_M}}^{\boldsymbol{\omega}}\right\|_{L^\infty} \leq \overline{\mathcal{B}}_D$.

Then, $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ are estimated by the following:

$$(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\omega}}) = \text{argmin}_{\boldsymbol{\psi}}\text{argmax}_{\boldsymbol{\omega}}\widehat{\mathcal{L}}_{\mathbf{IM}}(\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}, D_{\mathbf{I_M}}^{\boldsymbol{\omega}}; \widehat{\mathbf{G}}_{\mathbf{M}}), \tag{S2.10}$$

and the estimated conditional generator is $\widehat{\mathbf{I}}_{\mathbf{M}} = \mathbf{I}_{\mathbf{M}}^{\widehat{\boldsymbol{\psi}}}$ and the estimated discriminator is $\widehat{D}_{\mathbf{I_M}} = D_{\mathbf{I_M}}^{\widehat{\boldsymbol{\omega}}}$.

S2.5 $\mathbf{G_Y}$ estimation

We use two FNNs to estimate $\mathbf{G_Y}$ based on the empirical objective function $\widehat{\mathcal{L}}_{\mathbf{Y}}(\mathbf{G_Y}, D_{\mathbf{Y}})$. Let $\widetilde{\mathcal{Y}} := \mathcal{Y} \bigcup \mathcal{C}$. Following the approach in Section 3.1 in the paper, we denote the conditional generator network as $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$ parameterized by $\boldsymbol{\zeta}$, and the conditional discriminator network as $D_{\mathbf{Y}}^{\boldsymbol{\xi}}$ parameterized by $\boldsymbol{\xi}$.

• The generator network $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$: let $\mathcal{G}_Y \equiv \mathcal{G}_{\mathcal{H}_Y,\mathcal{W}_Y,\mathcal{S}_Y,\mathcal{B}_Y}$ be the set of ReLU neural networks $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \times \widetilde{\mathcal{Y}} \mapsto \mathcal{Y} \times \mathcal{Y}$ parameterized by $\boldsymbol{\zeta}$, depth $\mathcal{H}_Y$, width $\mathcal{W}_Y$, size $\mathcal{S}_Y$, and $\left\|G_{\mathbf{Y}}^{\boldsymbol{\zeta}}\right\|_{L^\infty} \leq \mathcal{B}_Y$.

• The discriminator network $D_{\mathbf{Y}}^{\boldsymbol{\xi}}$: denote $\mathcal{D}_Y \equiv \mathcal{D}_{\widetilde{\mathcal{H}}_Y,\widetilde{\mathcal{W}}_Y,\widetilde{\mathcal{S}}_Y,\widetilde{\mathcal{B}}_Y}$ as the set of ReLU neural networks $D_{\mathbf{Y}}^{\boldsymbol{\xi}} : \mathcal{X} \times \mathcal{M} \times \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, parameterized by $\boldsymbol{\xi}$, depth $\widetilde{\mathcal{H}}_Y$, width $\widetilde{\mathcal{W}}_Y$, size $\widetilde{\mathcal{S}}_Y$, and $\left\|D_{\mathbf{Y}}^{\boldsymbol{\xi}}\right\|_{L^\infty} \leq \widetilde{\mathcal{B}}_Y$.

Then, $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ are estimated by the following:

$$(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\xi}}) = \text{argmin}_{\boldsymbol{\zeta}} \text{argmax}_{\boldsymbol{\xi}} \widehat{\mathcal{L}}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}, D_{\mathbf{Y}}^{\boldsymbol{\xi}}), \tag{S2.11}$$

and the estimated conditional generator and discriminator are $\widehat{\mathbf{G}}_{\mathbf{Y}} = \mathbf{G}_{\mathbf{Y}}^{\widehat{\boldsymbol{\zeta}}}$ and $\widehat{D}_{\mathbf{Y}} = D_{\mathbf{Y}}^{\widehat{\boldsymbol{\xi}}}$, respectively.

S2.6 $\mathbf{I}_{\mathbf{Y}}$ estimation

We again use FNNs to estimate $\mathbf{I}_{\mathbf{Y}}$ based on the empirical objective function $\widehat{\mathcal{L}}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}}; \widehat{\mathbf{G}}_{\mathbf{Y}})$.

Denote the conditional generator network as $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$ parameterized by $\boldsymbol{\varphi}$, and the conditional discriminator network as $D_{\mathbf{I}_{\mathbf{Y}}}^{\boldsymbol{\lambda}}$ parameterized by $\boldsymbol{\lambda}$.

• The generator network $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$: denote $\mathcal{I}_Y \equiv \mathcal{I}_{\overline{\mathcal{H}}_Y,\overline{\mathcal{W}}_Y,\overline{\mathcal{S}}_Y,\overline{\mathcal{B}}_Y}$ as the set of ReLU neural networks $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}} : \mathcal{X} \times \mathcal{M} \mapsto \mathcal{Y} \times \mathcal{Y}$, parameterized by $\boldsymbol{\varphi}$, depth $\overline{\mathcal{H}}_Y$, width $\overline{\mathcal{W}}_Y$, size $\overline{\mathcal{S}}_Y$, and $\|\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}\|_{L^\infty} \leq \overline{\mathcal{B}}_Y$.

• The discriminator network $D_{\mathbf{I}_{\mathbf{Y}}}^{\boldsymbol{\lambda}}$: denote $\mathcal{D}_{I_Y} \equiv \mathcal{D}_{\overline{\mathcal{H}}_{D_Y},\overline{\mathcal{W}}_{D_Y},\overline{\mathcal{S}}_{D_Y},\overline{\mathcal{B}}_{D_Y}}$ as

the set of ReLU neural networks $D^{\boldsymbol{\lambda}}_{\mathbf{I_Y}} : \mathcal{X} \times \mathcal{M} \times \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, parameterized by $\boldsymbol{\lambda}$, depth $\overline{\mathcal{H}}_{D_Y}$, width $\overline{\mathcal{W}}_{D_Y}$, size $\overline{\mathcal{S}}_{D_Y}$, and $\left\|D^{\boldsymbol{\lambda}}_{\mathbf{I_Y}}\right\|_{L^\infty} \leq \overline{\mathcal{B}}_{D_Y}$. Then, $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$ are estimated by the following:

$$(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\lambda}}) = \text{argmin}_{\boldsymbol{\varphi}} \text{argmax}_{\boldsymbol{\lambda}} \widehat{\mathcal{L}}_{\mathbf{IY}}(\mathbf{I}^{\boldsymbol{\varphi}}_{\mathbf{Y}}, D^{\boldsymbol{\lambda}}_{\mathbf{I_Y}}; \widehat{\mathbf{G}}_{\mathbf{Y}}), \tag{S2.12}$$

and the estimated conditional generator and discriminator are $\widehat{\mathbf{I}}_{\mathbf{Y}} = \mathbf{I}^{\widehat{\boldsymbol{\varphi}}}_{\mathbf{Y}}$ and $\widehat{D}_{\mathbf{I_Y}} = D^{\widehat{\boldsymbol{\lambda}}}_{\mathbf{I_Y}}$, respectively.

S3 Regularity Conditions

(A.1) $\mathbf{G}^*_{\mathbf{M}}(\mathbf{z}, \mathbf{x}, 1, m)$ and $\mathbf{G}^*_{\mathbf{M}}(\mathbf{z}, \mathbf{x}, 0, m)$ are continuous in $(\mathbf{z}, \mathbf{x}, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M}$ with $\|\mathbf{G}^*_{\mathbf{M}}(\cdot, \cdot, 1, \cdot)\|_{L^\infty(\mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M})} \leq C_0$ and $\|\mathbf{G}^*_{\mathbf{M}}(\cdot, \cdot, 0, \cdot)\|_{L^\infty(\mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M})} \leq C_0$ for some constant $0 < C_0 < \infty$.

(A.2) For any $\mathbf{G}_{\mathbf{M}} \in \mathcal{G} \equiv \mathcal{G}_{\mathcal{H},\mathcal{W},\mathcal{S},\mathcal{B}}$,

$$\frac{p_{\mathbf{Q}}}{(p_{\mathbf{Q}} + p_{\mathbf{W}})} = \frac{p_{\mathbf{X}, G^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})), M_1(\mathbf{X})}}{p_{\mathbf{X}, G^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})), M_1(\mathbf{X})} + p_{\mathbf{X}, M_0(\mathbf{X}), G^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}} :$$

$\mathcal{X} \times \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is continuous and $0 < C_1 \leq p_{\mathbf{Q}}/(p_{\mathbf{Q}} + p_{\mathbf{W}}) \leq C_2 < 1$ for some constants $0 < C_1 \leq C_2 < \infty$.

(B.1) As sample size n goes to infinity, the network parameters of $\mathcal{G}$ satisfy

$$\mathcal{H}\mathcal{W} \to \infty \text{ and } \frac{\mathcal{B}\mathcal{S}\mathcal{H}\log(\mathcal{S})\log n}{n} \to 0. \tag{S3.13}$$

(B.2) As sample size n goes to infinity, the network parameters of $\mathcal{D}$ satisfy

$$\widetilde{\mathcal{H}}\widetilde{\mathcal{W}} \to \infty \text{ and } \frac{\widetilde{\mathcal{B}}\widetilde{\mathcal{S}}\widetilde{\mathcal{H}}\log(\widetilde{\mathcal{S}})\log n}{n} \to 0. \tag{S3.14}$$

(A.3) $\mathbf{I}^*_{\mathbf{M}}(\mathbf{z},\mathbf{x})$ is continuous in $(\mathbf{z},\mathbf{x}) \in \mathbb{R}^{d_z}\times\mathcal{X}$ with $\|\mathbf{I}^*_{\mathbf{M}}(\cdot,\cdot)\|_{L^\infty(\mathbb{R}^{d_z}\times\mathcal{X})} \leq C_7$ for some constant $0 < C_7 < \infty$.

(A.4) For any $\mathbf{I}_{\mathbf{M}} \in \mathcal{I} \equiv \mathcal{I}_{\overline{\mathcal{H}},\overline{\mathcal{W}},\overline{\mathcal{S}},\overline{\mathcal{B}}}$, $\frac{p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}}{p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}+p_{\mathbf{X},I^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X}),I^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})}}$ and $\frac{p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}}{p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}+p_{\mathbf{X},I^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X}),I^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})}}$ are continuous in $\mathcal{X}\times\mathcal{M}\times\mathcal{M}$, and they are bounded below by C_8 and are bounded above by C_9 for some constants $0 < C_8 \leq C_9 < \infty$.

(B.3) As sample size n goes to infinity, the network parameters of $\mathcal{I}$ satisfies

$$\overline{\mathcal{H}\mathcal{W}} \to \infty \text{ and } \frac{\overline{\mathcal{B}\mathcal{S}\mathcal{H}}\log(\overline{\mathcal{S}})\log n}{n} \to 0. \tag{S3.15}$$

(B.4) As sample size n goes to infinity, the network parameters of $\mathcal{D}_I$ satisfies

$$\overline{\mathcal{H}}_D\overline{\mathcal{W}}_D \to \infty \text{ and } \frac{\overline{\mathcal{B}}_D\overline{\mathcal{S}}_D\overline{\mathcal{H}}_D\log(\overline{\mathcal{S}}_D)\log n}{n} \to 0. \tag{S3.16}$$

(C.1) $\mathbf{G}^*_{\mathbf{Y}}(\widetilde{\mathbf{z}},\mathbf{x},1,m,y)$ and $\mathbf{G}^*_{\mathbf{Y}}(\widetilde{\mathbf{z}},\mathbf{x},0,m,y)$ are continuous in $(\widetilde{\mathbf{z}},\mathbf{x},m,y) \in \mathbb{R}^{d_z}\times\mathcal{X}\times\mathcal{M}\times\mathcal{Y}$ with $\|\mathbf{G}^*_{\mathbf{Y}}(\cdot,\cdot,1,\cdot,\cdot)\|_{L^\infty(\mathbb{R}^{d_z}\times\mathcal{X}\times\mathcal{M}\times\mathcal{Y})} \leq C_{Y0}$ and $\|\mathbf{G}^*_{\mathbf{Y}}(\cdot,\cdot,0,\cdot,\cdot)\|_{L^\infty(\mathbb{R}^{d_z}\times\mathcal{X}\times\mathcal{M}\times\mathcal{Y})} \leq C_{Y0}$ for some constant $0 < C_{Y0} < \infty$.

(C.2) For any $\mathbf{G}_{\mathbf{Y}} \in \mathcal{G}_Y \equiv \mathcal{G}_{\mathcal{H}_Y,\mathcal{W}_Y,\mathcal{S}_Y,\mathcal{B}_Y}$, $\frac{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\widetilde{\mathbf{z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)}}{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\widetilde{\mathbf{z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)}+p_{\mathbf{X},M,Y_0(\mathbf{X},M),G^{(1)}_{\mathbf{Y}}(\widetilde{\mathbf{z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))}} : \mathcal{X}\times\mathcal{M}\times\mathcal{Y}\times$

$\mathcal{Y} \to \mathbb{R}$ is continuous and it is not less than C_{Y1} and not larger than C_{Y2} for some constants $0 < C_{Y1} \leq C_{Y2} < \infty$.

(D.1) As sample size n_1 goes to infinity, the network parameters of $\mathcal{G}_Y$ satisfy

$$\mathcal{H}_Y \mathcal{W}_Y \to \infty \text{ and } \frac{\mathcal{B}_Y \mathcal{S}_Y \mathcal{H}_Y \log(\mathcal{S}_Y) \log n_1}{n_1} \to 0. \tag{S3.17}$$

(D.2) As sample size n_1 goes to infinity, the network parameters of $\mathcal{D}_Y$ satisfy

$$\widetilde{\mathcal{H}}_Y \widetilde{\mathcal{W}}_Y \to \infty \text{ and } \frac{\widetilde{\mathcal{B}}_Y \widetilde{\mathcal{S}}_Y \widetilde{\mathcal{H}}_Y \log(\widetilde{\mathcal{S}}_Y) \log n_1}{n_1} \to 0. \tag{S3.18}$$

And suppose that, for any given $\widehat{\mathbf{G}}_{\mathbf{Y}} \in \mathcal{G}_Y$,

(C.3) $\mathbf{I}^*_{\mathbf{Y}}(\mathbf{z}, \mathbf{x}, m)$ is continuous in $(\mathbf{z}, \mathbf{x}, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M}$ with $\|\mathbf{I}^*_{\mathbf{Y}}(\cdot, \cdot, \cdot)\|_{L^\infty(\mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M})} \leq C_{Y3}$ for some constant $0 < C_{Y3} < \infty$.

(C.4) For any $\mathbf{I}_{\mathbf{Y}} \in \mathcal{I}_Y \equiv \mathcal{I}_{\overline{\mathcal{H}}_Y, \overline{\mathcal{W}}_Y, \overline{\mathcal{S}}_Y, \overline{\mathcal{B}}_Y}$,

$$\frac{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)}}{p_{\mathbf{X},M,G^{(0)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)} + p_{\mathbf{X},M,I^{(0)}_{\mathbf{Y}}(\overline{\mathbf{Z}},\mathbf{X},M),I^{(1)}_{\mathbf{Y}}(\overline{\mathbf{Z}},\mathbf{X},M)}} \text{ and}$$

$$\frac{p_{\mathbf{X},M,Y_0(\mathbf{X},M),G^{(1)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))}}{p_{\mathbf{X},M,Y_0(\mathbf{X},M),G^{(1)}_{\mathbf{Y}}(\tilde{\mathbf{Z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))} + p_{\mathbf{X},M,I^{(0)}_{\mathbf{Y}}(\overline{\mathbf{Z}},\mathbf{X},M),I^{(1)}_{\mathbf{Y}}(\overline{\mathbf{Z}},\mathbf{X},M)}} \text{ are continuous in}$$

$\mathcal{X} \times \mathcal{M} \times \mathcal{Y} \times \mathcal{Y}$, and they are bounded below by C_{Y4} and are bounded above by C_{Y5} for some constants $0 < C_{Y4} \leq C_{Y5} < \infty$.

(D.3) As sample size n_1 goes to infinity, the network parameters of $\mathcal{I}_Y$ satisfy

$$\overline{\mathcal{H}}_Y \overline{\mathcal{W}}_Y \to \infty \text{ and } \frac{\overline{\mathcal{B}}_Y \overline{\mathcal{S}}_Y \overline{\mathcal{H}}_Y \log(\overline{\mathcal{S}}_Y) \log n_1}{n_1} \to 0. \tag{S3.19}$$

(D.4) As sample size n goes to infinity, the network parameters of $\mathcal{D}_{I_Y}$ satisfy

$$\overline{\mathcal{H}}_{D_Y}\overline{\mathcal{W}}_{D_Y} \to \infty \text{ and } \frac{\overline{\mathcal{B}}_{D_Y}\overline{\mathcal{S}}_{D_Y}\overline{\mathcal{H}}_{D_Y}\log(\overline{\mathcal{S}}_{D_Y})\log n_1}{n_1} \to 0. \tag{S3.20}$$

S4 Proofs of Theoretical Results in Section 4

Mediator Layer: we shall first show the convergence of the total variation norm

$$\begin{aligned}&\left\| p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} \right\|_{L^1} \\ &= \int_{\mathcal{X}\times\mathcal{M}\times\mathcal{M}} |p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\mathbf{x},m_0,m_1) - p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}(\mathbf{x},m_0,m_1)| d\mathbf{x}dm_0dm_1\end{aligned} \tag{S4.21}$$

as sample size n tends to infinity, under some assumptions, which are similar to those of Zhou et al. (2022).

Theorem S.1. *Under the assumptions (A.1), (A.2), (B.1), and (B.2), then* $\mathbb{E}_{S_n^M}\|p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}\|^2_{L^1} \to 0$, *as* $n \to \infty$.

Let l_0 be the number of samples in the sample set S_n^{IM} where $t_i = 0$, and l_1 be the number of samples in the sample set S_n^{IM} where $t_i = 1$. It is evident that $l_0 + l_1 = n$. As $n \to \infty$, at least one of the following statements holds true: (i) $l_1 \to \infty$; (ii) $l_0 \to \infty$. By employing a similar but simpler argument compared to the proof of Theorem S.1, we can establish:

Theorem S.2. *Under the assumptions (A.3), (A.4), (B.3), and (B.4), then*

$$(i)\ \ \mathbb{E}_{S_n^{IM}}\|p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}\|_{L^1} \to 0, \quad as \ \ l_1 \to \infty,$$

$$(ii)\ \ \mathbb{E}_{S_n^{IM}}\|p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}\|_{L^1} \to 0, \quad as \ \ l_0 \to \infty.$$

By combining Theorems S.1 and S.2, we can derive the following theorem:

Theorem S.3. *Under the assumptions (A.1)–(A.4) and (B.1)–(B.4), then*

$$\mathbb{E}_{S_n^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^n}\|p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}\|_{L^1} \to 0, \quad as \ \ n \to \infty,$$

$$\mathbb{E}_{S_n^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^n}\|p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}\|_{L^1} \to 0, \quad as \ \ n \to \infty.$$

Finally, we can derive Theorem 1.

Outcome Layer: Since we assume that the censoring rate α_r is fixed and strictly less than 1, the size of the non-censoring dataset $S_{n_1}^{(1)}$ tends to infinity as the size of the dataset S_n approaches infinity. Consequently, when n is sufficiently large and thus n_1 is also sufficiently large, we can exclusively utilize the non-censoring dataset $S_{n_1}^{(1)}$ to train the outcome layer.

Theorem S.4. *Under the assumptions (C.1)–(C.4) and (D.1)–(D.4), the following statements hold true, as* $n_1 \to \infty$*,*

$$\mathbb{E}_{S_n^Y \cup \{\overline{\mathbf{z}}_i\}_{i=1}^n}\|p_{\mathbf{X},M,\widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}},\mathbf{X},T=1,M,Y_1(\mathbf{X},M)),Y_1(\mathbf{X},M)} - p_{\mathbf{X},M,\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}},\mathbf{X},M),\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}},\mathbf{X},M)}\|_{L^1} \to 0,$$

$$\mathbb{E}_{S_n^Y \cup \{\overline{\mathbf{z}}_i\}_{i=1}^n}\|p_{\mathbf{X},M,Y_0(\mathbf{X},M),\widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}},\mathbf{X},T=0,M,Y_0(\mathbf{X},M))} - p_{\mathbf{X},M,\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}},\mathbf{X},M),\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}},\mathbf{X},M)}\|_{L^1} \to 0.$$

By employing a similar argument to the proof of Theorem 1 and utilizing results of Theorem S.4, we can derive Theorem 2.

S4.1 Proof of Theorem S.1

Proof. To handle the discrete variable $t \in \{0,1\}$ in a continuous manner, we introduce the following continuous function that smoothly connects the two states $t=0$ and $t=1$:

$$\varphi : (t, \mathbf{m}, \widetilde{\mathbf{m}}) \in [0,1] \times \mathcal{M}^2 \times \mathcal{M}^2 \mapsto \varphi(t, \mathbf{m}, \widetilde{\mathbf{m}}) = \begin{cases} \mathbf{m}, \ t \in [0, 1/4] \\ (2t - 1/2)\widetilde{\mathbf{m}} + (3/2 - 2t)\mathbf{m}, \ t \in [1/4, 3/4] \\ \widetilde{\mathbf{m}}, \ t \in [3/4, 1] \end{cases} \in \mathcal{M}^2, \tag{S4.22}$$

and then define the extended generator (to $t \in [0,1]$) as follows:

$$\widetilde{\mathbf{G}}^*_{\mathbf{M}} : (\mathbf{z}, \mathbf{x}, t, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times [0,1] \times \mathcal{M} \mapsto \varphi\Big(t, \mathbf{G}^*_{\mathbf{M}}(\mathbf{z}, \mathbf{x}, 0, m), \mathbf{G}^*_{\mathbf{M}}(\mathbf{z}, \mathbf{x}, 1, m)\Big) \in \mathcal{M}^2. \tag{S4.23}$$

Then, $\widetilde{\mathbf{G}}^*_{\mathbf{M}}$ is continuous with $\left\|\widetilde{\mathbf{G}}^*_{\mathbf{M}}\right\|_{L^\infty(\mathbb{R}^{d_z} \times \mathcal{X} \times [0,1] \times \mathcal{M})} \leq C_0$.

By a truncation argument, we only need to consider on domains $\Omega_1 = [-B, B]^{d_x + d_z} \times [0,1] \times [-B, B]$ and $\Omega_2 = [-B, B]^{d_x + 2}$ with $B = \log n$.

We note that

$$\begin{aligned}
&\mathbb{D}_{JS}(p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})}, p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))}) \\
&= \mathbb{D}_{KL}\left(p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} \middle\| \frac{p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} + p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))}}{2}\right) \\
&\quad + \mathbb{D}_{KL}\left(p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))} \middle\| \frac{p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} + p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))}}{2}\right) \\
&\geq \frac{1}{2}\left\| p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} - \frac{p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} + p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))}}{2} \right\|^2_{L_1} \\
&\quad + \frac{1}{2}\left\| p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))} - \frac{p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} + p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))}}{2} \right\|^2_{L_1} \\
&= \frac{1}{4}\| p_{\mathbf{X}, \widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} - p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))} \|^2_{L_1},
\end{aligned} \tag{S4.24}$$

where the inequality in the fourth line follows Pinskers inequalities (Tsybakov, 2008). Therefore,

$$\begin{aligned}
&\|p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}\|^2_{L_1} \\
\leq\ & 4\mathbb{D}_{JS}(p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}, p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}) \\
=\ & 4(\mathbb{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}) + \log 4) \\
=\ & 4(\mathbb{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}) - \mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}})), \qquad \text{(S4.25)}
\end{aligned}$$

where the equality in the last line follows the fact that $\mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}) = -\log 4$ from Lemma 1 and Equation (S2.5). So it suffices to show that the last line in (S4.25) converges to zero in expectation.

We follow the progress of proof in Zhou et al. (2022) and write

$$\begin{aligned}
0 \leq \mathbb{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}) - \mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}) =\ & \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D_{\mathbf{M}}) \\
=\ & \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}\in\mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&+ \sup_{D_{\mathbf{M}}\in\mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}\in\mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&+ \sup_{D_{\mathbf{M}}\in\mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}\in\mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&+ \sup_{D_{\mathbf{M}}\in\mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}\in\mathcal{D}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&+ \sup_{D_{\mathbf{M}}\in\mathcal{D}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&+ \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D_{\mathbf{M}})
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&\quad + \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&\quad + \sup_{D_{\mathbf{M}} \in \mathcal{D}} \widetilde{\mathcal{L}}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&\quad + \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D_{\mathbf{M}}), \\
&\leq \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \\
&\quad + 2 \sup_{D_{\mathbf{M}} \in \mathcal{D}, \mathbf{G}_{\mathbf{M}} \in \mathcal{G}} |\mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}) - \widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})| \\
&\quad + \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D_{\mathbf{M}}),
\end{aligned}
$$

where $\overline{\mathbf{G}}_{\mathbf{M}}$ is any element that belongs to $\mathcal{G}$. We then take infimum with respect to $\overline{\mathbf{G}}_{\mathbf{M}} \in \mathcal{G}$ on both sides of the above inequality and get

$$
\begin{aligned}
\mathbb{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}) - \mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}) \leq\ & \underbrace{\sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}})}_{\Delta_1} \\
& + \underbrace{2 \sup_{D_{\mathbf{M}} \in \mathcal{D}, \mathbf{G}_{\mathbf{M}} \in \mathcal{G}} |\mathcal{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}}) - \widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}, D_{\mathbf{M}})|}_{\Delta_2} \\
& + \underbrace{\inf_{\overline{\mathbf{G}}_{\mathbf{M}} \in \mathcal{G}} [\mathbb{L}_{\mathbf{M}}(\overline{\mathbf{G}}_{\mathbf{M}}) - \mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}})]}_{\Delta_3} \\
=\ & \Delta_1 + \Delta_2 + \Delta_3, \qquad \text{(S4.26)}
\end{aligned}
$$

where Δ_1 and Δ_3 are the approximation errors of $\mathcal{D}$ and $\mathcal{G}$ for their optimal counterparts, respectively, and Δ_2 is the statistical error and thus can be

further controlled by the empirical process theorem.

Lemma 5. *(Theorem 4.2 in Shen et al. (2019), Lemma B.5 in Zhou et al. (2022)) Let f be a uniformly continuous function defined on $\Omega \subset [-R, R]^d$. For arbitrary $L \in \mathbb{N}^+$ and $N \in \mathbb{N}^+$, there exists a function **ReLU** network f_ϕ with width $3^{d+3}\max\{d\lfloor N^{1/d}\rfloor, N+1\}$ and depth $12L+14+2d$ such that*

$$\|f - f_\phi\|_{L^\infty(\Omega)} \leq 19\sqrt{d}\omega_f^\Omega(2RN^{-2/d}L^{-2/d}), \tag{S4.27}$$

where $\lfloor\cdot\rfloor$ is the floor function and $\omega_f^\Omega(t)$ is the modulus of continuity of f satisfying $\omega_f^\Omega(t) \to 0$ as $t \to 0^+$.

Lemma 6. *(Lemma B.4 in Zhou et al. (2022)) If $\xi_i, i = 1, \ldots m$, are m finite linear combinations of Rademacher variables $\epsilon_j, j = 1, \ldots J$. Then*

$$\mathbb{E}_{\epsilon_j, j=1,\ldots J} \max_{1\leq i\leq m} |\xi_i| \leq C_4(\log m)^{1/2} \max_{1\leq i\leq m} \left(\mathbb{E}\xi_i^2\right)^{1/2}$$

for some constant $C_4 > 0$.

Lemma 7. *Under the assumptions* **(A.1)**, **(A.2)**, **(B.1)** *and* **(B.2)**, *the following statement is valid:*

$$\Delta_3 \equiv \inf_{\overline{\mathbf{G}}_\mathbf{M}\in\mathcal{G}} [\mathbb{L}_\mathbf{M}(\overline{\mathbf{G}}_\mathbf{M}) - \mathbb{L}_\mathbf{M}(\mathbf{G}_\mathbf{M}^*)] = o(1), \quad as \;\; n \to \infty.$$

Proof. By Assumption (**A.1**), $\widetilde{\mathbf{G}}_\mathbf{M}^*$ is continuous on $\Omega_1 = [-B, B]^{d_x+d_z} \times [0, 1] \times [-B, B]$ with $B = \log n$, and $\left\|\widetilde{\mathbf{G}}_\mathbf{M}^*\right\|_{L^\infty(\Omega_1)} \leq C_0$. Setting $L = \log n$,

$N = n^{\frac{d_x+d_z+2}{2(d_x+d_z+4)}}/\log n$, $\Omega = \Omega_1$ and $R = B$, in Lemma 5, we get an **ReLU** network $\overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}} \in \mathcal{G}$ with

$$\begin{aligned}
&\text{depth } \mathcal{H} = 12\log n + 14 + 2(d_x + d_z + 2),\\
&\text{width } \mathcal{W} = 3^{d_x+d_z+5}\max\{(d_x + d_z + 2)(n^{\frac{d_x+d_z+2}{2(d_x+d_z+4)}}/\log n)^{1/(d_x+d_z+2)}, n^{\frac{d_x+d_z+2}{2(d_x+d_z+4)}}/\log n + 1\},\\
&\text{size } \mathcal{S} = n^{\frac{d_x+d_z}{d_x+d_z+4}}/(\log^4 n) \text{ and } \mathcal{B} = 2C_0 \qquad \text{(S4.28)}
\end{aligned}$$

such that

$$\|\widetilde{\mathbf{G}}_{\mathbf{M}}^{*} - \overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}}\|_{L^\infty(\Omega_1)} \le 19\sqrt{d_x + d_z + 2}\,\omega_f^{\Omega_1}(2(\log n)n^{-1/(d_x+d_z+4)}). \quad \text{(S4.29)}$$

Thus, there exist at least one $t_0 \in [0, 1/4]$ and one $t_1 \in [3/4, 1]$ such that

$$\begin{aligned}
&\|\widetilde{\mathbf{G}}_{\mathbf{M}}^{*}(\cdot,\cdot,t_0,\cdot) - \overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}}(\cdot,\cdot,t_0,\cdot)\|_{L^\infty([-B,B]^{d_x+d_z+1})} \le 19\sqrt{d_x + d_z + 2}\,\omega_f^{\Omega_1}(2(\log n)n^{-1/(d_x+d_z+4)}),\\
&\|\widetilde{\mathbf{G}}_{\mathbf{M}}^{*}(\cdot,\cdot,t_1,\cdot) - \overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}}(\cdot,\cdot,t_1,\cdot)\|_{L^\infty([-B,B]^{d_x+d_z+1})} \le 19\sqrt{d_x + d_z + 2}\,\omega_f^{\Omega_1}(2(\log n)n^{-1/(d_x+d_z+4)}),
\end{aligned}$$

that is,

$$\begin{aligned}
&\|\mathbf{G}_{\mathbf{M}}^{*}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X})) - \overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}}(\cdot,\cdot,t_0,\cdot)\|_{L^\infty([-B,B]^{d_x+d_z+1})} \to 0, \quad \text{as } n \to \infty,\\
&\|\mathbf{G}_{\mathbf{M}}^{*}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})) - \overline{\mathbf{G}}_{\mathbf{M}}^{\boldsymbol{\theta}}(\cdot,\cdot,t_1,\cdot)\|_{L^\infty([-B,B]^{d_x+d_z+1})} \to 0, \quad \text{as } n \to \infty.
\end{aligned} \qquad \text{(S4.30)}$$

Let

$$\overline{D}_{\mathbf{M}}(\eta) = \frac{p_{\mathbf{X},\overline{G}_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{Z},\mathbf{X},t_1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta)}{p_{\mathbf{X},\overline{G}_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{Z},\mathbf{X},t_1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta) + p_{\mathbf{X},M_0(\mathbf{X}),\overline{G}_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{Z},\mathbf{X},t_0,M_0(\mathbf{X}))}(\eta)}, \qquad \text{(S4.31)}$$

$$D^*_{\mathbf{M}}(\eta) = \frac{p_{\mathbf{X},G^{*,(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta)}{p_{\mathbf{X},G^{*,(0)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta) + p_{\mathbf{X},M_0(\mathbf{X}),G^{*,(1)}_{\mathbf{M}}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}(\eta)}, \tag{S4.32}$$

where $\eta \in X \times \mathcal{M} \times \mathcal{M}$. Then, from the display on (S4.30) and continuity, we have

$$\|D^*_{\mathbf{M}} - \overline{D}_{\mathbf{M}}\|_{L^\infty([-B,B]^{d_x+d_z+1})} \to 0 \text{ as } n \to \infty, \tag{S4.33}$$

and therefore, we obtain

$$\mathbb{L}_{\mathbf{M}}(\overline{\mathbf{G}}^{\boldsymbol{\theta}}_{\mathbf{M}}) = \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}^{\boldsymbol{\theta}}_{\mathbf{M}}, D_{\mathbf{M}}) = \mathcal{L}_{\mathbf{M}}(\overline{\mathbf{G}}^{\boldsymbol{\theta}}_{\mathbf{M}}, \overline{D}_{\mathbf{M}})$$

converge to

$$\mathbb{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}) = \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D_{\mathbf{M}}) = \mathcal{L}_{\mathbf{M}}(\mathbf{G}^*_{\mathbf{M}}, D^*_{\mathbf{M}})$$

as $n \to \infty$. □

Lemma 8. *Under the assumptions* **(A.1)**, **(A.2)**, **(B.1)** *and* **(B.2)**, *the following statement is valid:*

$$\Delta_2 \leq \mathcal{O}(n^{-\frac{2}{4+d_x+d_z}} + n^{-\frac{2}{4+d_x}}). \tag{S4.34}$$

Proof. To bound the statistical error Δ_2, we follow the empirical process argument in lemma B.2 in Zhou et al. (2022). By Assumption (**A.2**), for

any $\mathbf{G_M} \in \mathcal{G}$,

$$D_{\mathbf{M}}(\cdot) = \frac{p_{\mathbf{X},G_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\cdot)}{p_{\mathbf{X},G_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\cdot) + p_{\mathbf{X},M_0(\mathbf{X}),G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}(\cdot)} : \mathcal{X} \times \mathcal{M} \times \mathcal{M} \to \mathbb{R} \tag{S4.35}$$

is continuous and $0 < C_1 \leq \inf_{\eta\in\Omega_2} D_{\mathbf{M}}(\eta) \leq \sup_{\eta\in\Omega_2} D_{\mathbf{M}}(\eta) \leq C_2 < 1$.

Setting $L = \log n$, $N = n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n$, $\Omega = \Omega_2$ and $R = B$, in Lemma 5,

we get an **ReLU** network $\overline{D}_{\mathbf{M}}^{\overline{\phi}} \in \mathcal{D}$ with

$$\begin{aligned}
&\text{depth}\widetilde{\mathcal{H}} = 12\log n + 14 + 2(d_x+2),\\
&\text{width } \widetilde{\mathcal{W}} = 3^{d_x+5}\max\{(d_x+2)(n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n)^{1/(d_x+2)}, n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n + 1\},\\
&\text{size } \widetilde{\mathcal{S}} = n^{\frac{d_x}{d_x+4}}/(\log^4 n) \text{ and } \widetilde{\mathcal{B}} = 2C_2
\end{aligned} \tag{S4.36}$$

such that

$$\|D_{\mathbf{M}} - \overline{D}_{\mathbf{M}}^{\overline{\phi}}\|_{L^\infty(\Omega_2)} \leq 19\sqrt{d_x+2}\,\omega_f^{\Omega_2}(2(\log n)n^{-1/(d_x+4)}). \tag{S4.37}$$

Let $(\mathbf{X}, T, M) \sim P_{\mathbf{X},T,M}$ and $(\mathbf{x}_i, t_i, m_i)$, $i = 1, ..., n$ are i.i.d copies of $(\mathbf{X}, T, M)$. Let $\mathbf{Z} \sim P_{\mathbf{Z}}$ and $\mathbf{Z} \perp\!\!\!\perp (\mathbf{X}, T, M)$, $\mathbf{z}_j$, $j = 1, ..., n$ are i.i.d copies of $\mathbf{Z}$. Then, $\mathbf{s}_i = (\mathbf{x}_i, t_i, m_i, \mathbf{z}_i)$ are i.i.d copies of $\mathbf{S} = (\mathbf{X}, T, M, \mathbf{Z}) \sim P_{\mathbf{X},T,M}P_{\mathbf{Z}}$. Denote

$$\begin{aligned}
b(\mathbf{G_M}, D_{\mathbf{M}}; \mathbf{S}) =& T\log D_{\mathbf{M}}(\mathbf{X}, (1-T)M + tG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M))\\
&+ (1-T)\log[1 - D_{\mathbf{M}}(\mathbf{X}, (1-T)M + TG_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T,M), TM + (1-T)G_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T,M))].
\end{aligned} \tag{S4.38}$$

Then, $\mathcal{L}_{\mathbf{M}}(\mathbf{G_M}, D_{\mathbf{M}}) = \mathbb{E}_{\mathbf{S}}[b(\mathbf{G_M}, D_{\mathbf{M}}; \mathbf{S})]$ and $\widetilde{\mathcal{L}}_{\mathbf{M}}(\mathbf{G_M}, D_{\mathbf{M}}) = \frac{1}{n}\sum_{i=1}^n b(\mathbf{G_M}, D_{\mathbf{M}}; \mathbf{s}_i)]$.

Let ϵ_i, $i = 1, ..., n$ be i.i.d Rademacher random samples that are independent

of $\mathbf{s}_i$, $i = 1, ..., n$. Denote the Rademacher complexity of $\mathcal{D} \times \mathcal{G}$ (Bartlett and Mendelson, 2002) by

$$\mathcal{C}(\mathcal{D} \times \mathcal{G}) = \frac{1}{n}\mathbb{E}_{\{\mathbf{s}_i, \epsilon_i\}_{i=1}^n}\left[\sup_{\mathbf{G_M} \in \mathcal{G}, D_\mathbf{M} \in \mathcal{D}} \left| \sum_{i=1}^n \epsilon_i b\left(\mathbf{G_M}, D_\mathbf{M}; \mathbf{s}_i\right) \right|\right]. \quad \text{(S4.39)}$$

Let $\mathfrak{C}\left(\mathcal{D} \times \mathcal{G}, e_{n,1}, \delta\right))$ be the covering number of $\mathcal{D} \times \mathcal{G}$ with respect to the empirical distance

$$e_{n,1}((\mathbf{G_M}, D_\mathbf{M}), (\widetilde{\mathbf{G}}_\mathbf{M}, \widetilde{D}_\mathbf{M})) := \frac{1}{n}\mathbb{E}_{\epsilon_i}\left[\sum_{i=1}^n \left|\epsilon_i \left(b\left(\mathbf{G_M}, D_\mathbf{M}; \mathbf{s}_i\right) - b\left(\widetilde{\mathbf{G}}_\mathbf{M}, \widetilde{D}_\mathbf{M}; \mathbf{s}_i\right)\right)\right|\right]. \quad \text{(S4.40)}$$

Also define

$$e_{n,\infty}((\mathbf{G_M}, D_\mathbf{M}), (\widetilde{\mathbf{G}}_\mathbf{M}, \widetilde{D}_\mathbf{M})) := \mathbb{E}_{\epsilon_i}\left[\sup_{1 \le i \le n} \left|\epsilon_i \left(b\left(\mathbf{G_M}, D_\mathbf{M}; \mathbf{s}_i\right) - b\left(\widetilde{\mathbf{G}}_\mathbf{M}, \widetilde{D}_\mathbf{M}; \mathbf{s}_i\right)\right)\right|\right]. \quad \text{(S4.41)}$$

First, by the standard symmetrization technique and the law of iterated expectations, we have

$$\begin{aligned} &\sup_{D_\mathbf{M} \in \mathcal{D}, \mathbf{G_M} \in \mathcal{G}} |\mathcal{L}_\mathbf{M}(\mathbf{G_M}, D_\mathbf{M}) - \widetilde{\mathcal{L}}_\mathbf{M}(\mathbf{G_M}, D_\mathbf{M})| = \ 2\mathcal{C}(\mathcal{D} \times \mathcal{G}) \\ &= \ 2\mathbb{E}_{\mathbf{s}_1, \ldots, \mathbf{s}_n}\left\{\mathbb{E}_{\epsilon_i, i=1,\ldots,n}\left[\mathcal{C}(\mathcal{G} \times \mathcal{D}) \mid (\mathbf{s}, \ldots, \mathbf{s}_n)\right]\right\}. \end{aligned} \quad \text{(S4.42)}$$

For $\delta > 0$, let $\mathcal{D}_\delta \times \mathcal{G}_\delta$ be such a covering set at scale δ of $\mathcal{D} \times \mathcal{G}$. Then, by the triangle inequality and Lemma 6, we have

$$\mathbb{E}_{\mathbf{s}_1, \ldots, \mathbf{s}_n}\left\{\mathbb{E}_{\epsilon_i, i=1,\ldots,n}\left[\mathcal{C}(\mathcal{G} \times \mathcal{D}) \mid (\mathbf{s}_1, \ldots, \mathbf{s}_n)\right]\right\}$$

$$\leq \delta + \frac{1}{n}\mathbb{E}_{\mathbf{s}_1,\ldots,\mathbf{s}_n}\left\{\mathbb{E}_{\epsilon_i,i=1,\ldots,n}\Big[\sup_{(\mathbf{G_M},D_\mathbf{M})\in\mathcal{D}_\delta\times\mathcal{G}_\delta}\Big|\sum_{i=1}^{n}\epsilon_i b\left(\mathbf{G_M},D_\mathbf{M};\mathbf{s}_i\right)\Big|\,|\,(\mathbf{s}_1,\ldots,\mathbf{s}_n)\Big]\right\}$$
$$\leq \delta + C_4\frac{1}{n}\mathbb{E}_{\mathbf{s}_1,\ldots,\mathbf{s}_n}\left\{\big[\log\mathfrak{C}\left(\mathcal{D}\times\mathcal{G},e_{n,1},\delta\right)\big]^{1/2}\max_{(\mathbf{G_M},D_\mathbf{M})\in\mathcal{D}_\delta\times\mathcal{G}_\delta}\big[\sum_{i=1}^{n}b^2\left(\mathbf{G_M},D_\mathbf{M};\mathbf{s}_i\right)\big]^{1/2}\right\}. \tag{S4.43}$$

Since $\|b(\mathbf{G_M},D_\mathbf{M};\mathbf{S})\|_{L^\infty}\leq C_3:=\max\{|\log C_1|,|\log(1-C_2)|\}$, we have

$$\left[\sum_{i=1}^{n}b^2\left(\mathbf{G_M},D_\mathbf{M};\mathbf{s}_i\right)\right]^{1/2}\leq\sqrt{n}C_3.$$

Therefore,

$$\frac{1}{n}\mathbb{E}_{\mathbf{s}_1,\ldots,\mathbf{s}_n}\left\{\left(\log\mathfrak{C}\left(\mathcal{D}\times\mathcal{G},e_{n,1},\delta\right)\right)^{1/2}\max_{(\mathbf{G_M},D_\mathbf{M})\in\mathcal{D}_\delta\times\mathcal{G}_\delta}\left[\sum_{i=1}^{n}b^2\left(\mathbf{G_M},D_\mathbf{M};\mathbf{s}_i\right)\right]^{1/2}\right\}$$
$$\leq\frac{1}{n}\mathbb{E}_{\mathbf{s}_1,\ldots,\mathbf{s}_n}\left[\left(\log\mathfrak{C}\left(\mathcal{D}\times\mathcal{G},e_{n,1},\delta\right)\right)^{1/2}\sqrt{n}C_3\right]$$
$$\leq\frac{C_3}{\sqrt{n}}\left[\log\mathfrak{C}\left(\mathcal{D},e_{n,1},\delta\right)+\log\mathfrak{C}\left(\mathcal{G},e_{n,1},\delta\right)\right]^{1/2}. \tag{S4.44}$$

Now since $\mathfrak{C}\left(\mathcal{G},e_{n,1},\delta\right)\leq\mathfrak{C}\left(\mathcal{G},e_{n,\infty},\delta\right)$ (similar result for $\mathcal{D}$) and

$$\log\mathfrak{C}\left(\mathcal{G},e_{n,\infty},\delta\right))\leq\mathrm{Pdim}_\mathcal{G}\log\frac{2e\mathcal{B}n}{\delta\,\mathrm{Pim}_\mathcal{G}},$$

where $\mathrm{Pdim}_\mathcal{G}$ is the Pseudo dimension of $\mathcal{G}_{\mathcal{H},\mathcal{W},\mathcal{S},\mathcal{B}}$, which satisfies Bartlett et al. (2019)

$$C_5\mathcal{H}\mathcal{S}\log\mathcal{S}\leq\mathrm{Pim}_\mathcal{G}\leq C_6\mathcal{H}\mathcal{S}\log\mathcal{S},$$

for some positive constants C_5 and C_6. Then, we have

$$\frac{1}{\sqrt{n}}\left[\log\mathfrak{C}\left(\mathcal{D},e_{n,1},\delta\right)+\log\mathfrak{C}\left(\mathcal{G},e_{n,1},\delta\right)\right]^{1/2}$$

$$\lesssim \frac{1}{\sqrt{n}}\left[\mathcal{H}\mathcal{S}\log\mathcal{S}\log\frac{\mathcal{B}n}{\delta\mathcal{H}\mathcal{S}\log\mathcal{S}}+\widetilde{\mathcal{H}}\widetilde{\mathcal{S}}\log\widetilde{\mathcal{S}}\log\frac{\widetilde{\mathcal{B}}n}{\delta\widetilde{\mathcal{H}}\widetilde{\mathcal{S}}\log\widetilde{\mathcal{S}}}\right]^{1/2}. \quad \text{(S4.45)}$$

As a result, (S4.34) follows from (S4.28), (S4.36) and (S4.42)–(S4.45) with the selection of the network parameters of $\mathcal{D}_{\widetilde{\mathcal{H}},\widetilde{\mathcal{W}},\widetilde{\mathcal{S}},\widetilde{\mathcal{B}}}, \mathcal{G}_{\mathcal{H},\mathcal{W},\mathcal{S},\mathcal{B}}$ and with $\delta = \frac{1}{n}$. □

Lemma 9. *Under the assumptions* **(A.1)**, **(A.2)**, **(B.1)** *and* **(B.2)**, *the following statement is valid:*

$$\mathbb{E}_{S_n^M}[\Delta_1] \equiv \mathbb{E}_{S_n^M}[\sup_{D_{\mathbf{M}}}\mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}}\in\mathcal{D}}\mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}})] \to 0 \quad as \quad n\to\infty. \quad \text{(S4.46)}$$

Proof. Conditioning on the data S_n^M, $\sup_{D_{\mathbf{M}}}\mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}})$ is attained at

$$D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}}(\eta) = \frac{p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta)}{p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(\eta) + p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))}(\eta)}.$$

By Assumption (**A.2**), $D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}}(\eta) : \mathcal{X}\times\mathcal{M}\times\mathcal{M}\to\mathbb{R}$ is continuous on $\Omega_2 = [-B,B]^{d_x+2}$ with $B = \log n$, and $\|D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}}\|_{L^\infty(\Omega_2)} \le C_2$. Setting $L = \log n$, $N = n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n$, $\Omega = \Omega_2$ and $R = B$, in Lemma 5, we get an **ReLU** network $\widehat{D}_{\mathbf{M}}^{\boldsymbol{\phi}} \in \mathcal{D}$ with depth $\widetilde{\mathcal{H}} = 12\log n + 14 + 2(d_x+2)$, $\widetilde{\mathcal{W}} = 3^{d_x+5}\max\{(d_x+2)(n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n)^{1/(d_x+2)}, n^{\frac{d_x+2}{2(2+d_x+2)}}/\log n + 1\}$, and size $\widetilde{\mathcal{S}} = n^{\frac{d_x}{d_x+4}}/(\log^4 n)$, $\widetilde{\mathcal{B}} = 2C_2$ such that

$$\|D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}} - \widehat{D}_{\mathbf{M}}^{\boldsymbol{\phi}}\|_{L^\infty(\Omega_2)} \le 19\sqrt{d_x+2}\,\omega_f^{\Omega_2}(2(\log n)n^{-1/(d_x+4)}), \quad \text{(S4.47)}$$

this is $\|D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}} - \widehat{D}_{\mathbf{M}}^{\phi}\|_{L^\infty(\Omega_2)} \to 0$, as $n \to \infty$. Then

$$0 < \sup_{D_{\mathbf{M}}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) - \sup_{D_{\mathbf{M}} \in \mathcal{D}} \mathcal{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}) \leq \mathcal{L}_1(\widehat{\mathbf{G}}_{\mathbf{M}}, D_{\mathbf{M}}^{\widehat{\mathbf{G}}_{\mathbf{M}}}) - \mathcal{L}_1(\widehat{\mathbf{G}}_{\mathbf{M}}, \widehat{D}_{\mathbf{M}}^{\phi}) \to 0,$$

by continuity. □

Thus, combined with the results of the above three lemmas, we can derive the result of Theorem S.1:

$$\begin{aligned}
&\mathbb{E}_{S_n^M} \| p_{\mathbf{X}, \widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} - p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))} \|_{L_1}^2 \\
&\leq \mathbb{E}_{S_n^M} 4(\mathbb{L}_{\mathbf{M}}(\widehat{\mathbf{G}}_{\mathbf{M}}) - \mathbb{L}_{\mathbf{M}}(\mathbf{G}_{\mathbf{M}}^*)) \\
&\leq 4\mathbb{E}_{S_n^M} \Delta_1 + \Delta_2 + \Delta_3 \to 0, \quad \text{as} \quad n \to \infty.
\end{aligned}$$

□

S4.2 Proof of Theorem S.3

Proof. Without loss of generality, we assume that $l_1 \to \infty$ as $n \to \infty$. Otherwise, if $l_0 \to \infty$ as $n \to \infty$, we can employ a similar argument as follows. For an arbitrary $\epsilon > 0$, by Theorem S.1, there exists $N_0 > 0$ such that, for any $n \geq N_0$,

$$\mathbb{E}_{S_n^M} \| p_{\mathbf{X}, \widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z}, \mathbf{X}, T=1, M_1(\mathbf{X})), M_1(\mathbf{X})} - p_{\mathbf{X}, M_0(\mathbf{X}), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{Z}, \mathbf{X}, T=0, M_0(\mathbf{X}))} \|_{L^1} \leq \frac{1}{2}\epsilon. \tag{S4.48}$$

Let $\widehat{\mathbf{G}}_{\mathbf{M},N_0}$ be the estimated conditional generator corresponding to the sample set $S_{N_0}^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^{N_0}$. Then,

$$\mathbb{E}_{S_{N_0}^M} \| p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M},N_0}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M},N_0}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} \|_{L^1} < \frac{1}{2}\epsilon. \tag{S4.49}$$

Next, for any $n \geq N_0$, we can define the sample set $S_n^{IM} := \{(\mathbf{x}_i, t_i, \overline{m}_i^{(0)}, \overline{m}_i^{(1)}, \widehat{\mathbf{z}}_i)\}_{i=1}^n$, where $(\overline{m}_i^{(0)}, \overline{m}_i^{(1)}) = t_i \cdot (\widehat{G}^{(0)}_{\mathbf{M},N_0}(\mathbf{z}_i, \mathbf{x}_i, T=1, m_i), m_i) + (1-t_i)\cdot(m_i, \widehat{G}^{(1)}_{\mathbf{M},N_0}(\mathbf{z}_i, \mathbf{x}_i, T=0, m_i))$, and thus, by Theorem S.2, there exists $N_1 > N_0$ such that, for any $n \geq N_1$,

$$\mathbb{E}_{S_n^{IM}} \| p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M},N_0}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1} < \frac{1}{2}\epsilon, \tag{S4.50}$$

which implies

$$\mathbb{E}_{S_n^M} \| p_{\mathbf{X},\widehat{G}^{(0)}_{\mathbf{M},N_0}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1} < \frac{1}{2}\epsilon. \tag{S4.51}$$

In addition, by (S4.49) and (S4.50),

$$\mathbb{E}_{S_n^M} \| p_{\mathbf{X},M_0(\mathbf{X}),\widehat{G}^{(1)}_{\mathbf{M},N_0}(\mathbf{Z},\mathbf{X},T=0,M_0(\mathbf{X}))} - p_{\mathbf{X},\widehat{I}^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1} < \epsilon. \tag{S4.52}$$

Therefore, by (S4.51), (S4.52) and the arbitrary nature of ϵ, we have the results of Theorem S.3. □

S4.3 Proof of Theorem 1

Proof. By using the first result of Theorem S.3, we have

$$\mathbb{E}_{S_n^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^n} \| p_{\mathbf{X},M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1}$$

$$=\mathbb{E}_{S_n^M\cup\{\widehat{\mathbf{z}}_i\}_{i=1}^n}\int_{\mathcal{X}\times\mathcal{M}}\left|p_{\mathbf{X},M_1(\mathbf{X})}(x,m)-p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}(x,m)\right|dmdx$$

$$=\mathbb{E}_{S_n^M\cup\{\widehat{\mathbf{z}}_i\}_{i=1}^n}\int_{\mathcal{X}\times\mathcal{M}}\Big|\int_{\mathcal{M}}p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(x,\widetilde{m},m)d\widetilde{m}-\int_{\mathcal{M}}p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}(x,\widetilde{m},m)d\widetilde{m}\Big|dmdx$$

$$\leq\mathbb{E}_{S_n^M\cup\{\widehat{\mathbf{z}}_i\}_{i=1}^n}\int_{\mathcal{X}\times\mathcal{M}\times\mathcal{M}}\left|p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}(x,\widetilde{m},m)-p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}(x,\widetilde{m},m)\right|d\widetilde{m}dmdx$$

$$=\mathbb{E}_{S_n^M\cup\{\widehat{\mathbf{z}}_i\}_{i=1}^n}\|p_{\mathbf{X},\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{Z},\mathbf{X},T=1,M_1(\mathbf{X})),M_1(\mathbf{X})}-p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}),\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})}\|_{L^1}\to 0.$$

By applying a similar augmentation technique as above and utilizing the second result of Theorem S.3, we can demonstrate the validity of the second result of this Theorem. □

S5 Implementation of CGAN-ICMA-SO

We describe the implementation of CGAN-ICMA-SO. We use ReLU as the activation function to train the generator $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}, \mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}, \mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}, \mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$ and the discriminator $D_{\mathbf{M}}^{\boldsymbol{\phi}}, D_{\mathbf{I}_{\mathbf{M}}}^{\boldsymbol{\omega}}, D_{\mathbf{Y}}^{\boldsymbol{\xi}}, D_{\mathbf{I}_{\mathbf{Y}}}^{\boldsymbol{\lambda}}$. We train the discriminator and the generator iteratively by updating $\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\zeta}, \boldsymbol{\varphi}, \boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ as follows:

(a) Fix $\boldsymbol{\theta}$, update the discriminator $D_{\mathbf{M}}^{\boldsymbol{\phi}}$ by ascending the stochastic gradient of the loss (S2.2) with respect to $\boldsymbol{\phi}$.

(b) Fix $\boldsymbol{\phi}$, update the generator $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$ by descending the stochastic gradient of the loss (S2.2) with respect to $\boldsymbol{\theta}$.

(c) Fix $\boldsymbol{\zeta}$, update the discriminator $D_{\mathbf{Y}}^{\boldsymbol{\xi}}$ by ascending the stochastic gradient of the loss (3.10) with respect to $\boldsymbol{\xi}$.

(d) Fix $\boldsymbol{\xi}$, update the generator $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$ by descending the stochastic gradient of the loss (3.10) with respect to $\boldsymbol{\zeta}$.

(e) Fix $\boldsymbol{\psi}$, update the discriminator $D_{\mathbf{I}_{\mathbf{M}}}^{\boldsymbol{\omega}}$ by ascending the stochastic gradient of the loss (S2.3) with respect to $\boldsymbol{\omega}$.

(f) Fix $\boldsymbol{\omega}$, update the generator $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}$ by descending the stochastic gradient of the loss (S2.3) with respect to $\boldsymbol{\psi}$.

(g) Fix $\boldsymbol{\varphi}$, update the discriminator $D_{\mathbf{I}_{\mathbf{Y}}}^{\boldsymbol{\lambda}}$ by ascending the stochastic gradient of the loss (3.11) with respect to $\boldsymbol{\lambda}$.

(h) Fix $\boldsymbol{\lambda}$, update the generator $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$ by descending the stochastic gradient of the loss (3.11) with respect to $\boldsymbol{\varphi}$.

The training process is described below.

Algorithm 1 Training CGAN-ICMA-SO

Input: (a) Samples $\{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = \widetilde{y}_i, \delta = \delta_i\}_{i=1}^n = \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, Y = y_i, \delta = \delta_i = 1\}_{i=1}^{n_1} \cup \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, C = c_i, \delta = \delta_i = 0\}_{i=n_1+1}^n$; (b) Samples $\{\mathbf{Z} = \mathbf{z}_i\}_{i=1}^n$ from $P_{\mathbf{Z}}$; (c) Samples $\{\widetilde{\mathbf{Z}} = \widetilde{\mathbf{z}}_i\}_{i=1}^n$ from $P_{\widetilde{\mathbf{Z}}}$; (d) Samples $\{\widehat{\mathbf{Z}} = \widehat{\mathbf{z}}_i\}_{i=1}^n$ from $P_{\widehat{\mathbf{Z}}}$; (e) Samples $\{\overline{\mathbf{Z}} = \overline{\mathbf{z}}_i\}_{i=1}^n$ from $P_{\overline{\mathbf{Z}}}$

Output: Conditional generator $\widehat{\mathbf{G}}_{\mathbf{M}}, \widehat{\mathbf{G}}_{\mathbf{Y}}, \widehat{\mathbf{I}}_{\mathbf{M}}, \widehat{\mathbf{I}}_{\mathbf{Y}}$, and discriminator $\widehat{D}_{\mathbf{M}}$, $\widehat{D}_{\mathbf{Y}}, \widehat{D}_{\mathbf{I}_{\mathbf{M}}}, \widehat{D}_{\mathbf{I}_{\mathbf{Y}}}$

Optimizer: Adam (Kingma and Ba, 2014)

First Step:

while not converged **do**

- Compute $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i) = (G_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i), G_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i))$, $i = 1, \ldots, n$ and $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, \widetilde{y}_i) = (G_{\mathbf{Y}}^{\boldsymbol{\zeta},(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, \widetilde{y}_i), G_{\mathbf{Y}}^{\boldsymbol{\zeta},(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, \widetilde{y}_i)), i = 1, \ldots, n$, where $\{\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, \widetilde{y}_i), i = 1, \ldots, n\} = \{\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i), i = 1, \ldots, n_1\} \cup \{\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, c_i), i = n_1 + 1, \ldots, n\}$. Let $S_t = \{\mathbf{x}_i, t_i, m_i, \widetilde{y}_i, \mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i), \mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, \widetilde{y}_i), \delta_i, i = 1, \ldots, n\} = S_{t1} \cup S_{t2}$, where $S_{t1} = \{\mathbf{x}_i, t_i, m_i, y_i, \mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i), \mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i), \delta_i = 1, i = 1, \ldots, n_1\}$, and $S_{t2} = \{\mathbf{x}_i, t_i, m_i, c_i, \mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}(\mathbf{z}_i, \mathbf{x}_i, t_i, m_i), \mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, c_i), \delta_i = 0, i = n_1 + 1, \ldots, n\}$.

- Randomly select B samples from S_t, where B_1 samples from S_{t1}, B_2 samples from S_{t2}, and $B = B_1 + B_2$. Denote the subscripts of the selected samples by $\{b_i : i = 1, \ldots, B\}, \{b_i : i = 1, \ldots, B_1\}$, and $\{b_i : i = B_1 + 1, \ldots, B\}$.

- Update $D_{\mathbf{M}}^{\boldsymbol{\phi}}$ and $D_{\mathbf{Y}}^{\boldsymbol{\xi}}$ by ascending their stochastic gradients:

$$\nabla_{\boldsymbol{\phi}} \left\{ \frac{1}{B} \sum_{i=1}^{B} \left\{ t_{b_i} \log D_{\mathbf{M}}^{\boldsymbol{\phi}} \left(\mathbf{x}_{b_i}, (1 - t_{b_i}) m_{b_i} + t_{b_i} G_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}), t_{b_i} m_{b_i} + (1 - t_{b_i}) G_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}) \right) \right.\right.$$
$$\left.\left. + (1 - t_{b_i}) \log \left[1 - D_{\mathbf{M}}^{\boldsymbol{\phi}} \left(\mathbf{x}_{b_i}, (1 - t_{b_i}) m_{b_i} + t_{b_i} G_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}), t_{b_i} m_{b_i} + (1 - t_{b_i}) G_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}) \right) \right] \right\} \right\}$$

$$\nabla_{\boldsymbol{\xi}} \left\{ \frac{1}{B_1} \sum_{i=1}^{B_1} \left\{ t_{b_i} \log D_{\mathbf{Y}}^{\boldsymbol{\xi}} \left(\mathbf{x}_{b_i}, m_{b_i}, (1 - t_{b_i}) y_{b_i} + t_{b_i} G_{\mathbf{Y}}^{\boldsymbol{\zeta},(0)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}), t_{b_i} y_{b_i} + (1 - t_{b_i}) G_{\mathbf{Y}}^{\boldsymbol{\zeta},(1)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}) \right) \right.\right.$$
$$\left.\left. + (1 - t_{b_i}) \log \left[1 - D_{\mathbf{Y}}^{\boldsymbol{\xi}} \left(\mathbf{x}_{b_i}, m_{b_i}, (1 - t_{b_i}) y_{b_i} + t_{b_i} G_{\mathbf{Y}}^{\boldsymbol{\zeta},(0)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}), t_{b_i} y_{b_i} + (1 - t_{b_i}) G_{\mathbf{Y}}^{\boldsymbol{\zeta},(1)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}) \right) \right] \right\} \right\}$$

- Update $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$ and $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$ by descending their stochastic gradients:

$$\begin{aligned}
&\bigtriangledown_{\boldsymbol{\theta}} \Bigg\{ \frac{1}{B}\sum_{i=1}^{B} \Big\{ t_{b_i} \log D_{\mathbf{M}}^{\boldsymbol{\phi}}\Big(\mathbf{x}_{b_i}, (1-t_{b_i})m_{b_i} + t_{b_i} G_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}), t_{b_i} m_{b_i} + (1-t_{b_i}) G_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i})\Big) \\
&\quad + (1-t_{b_i}) \log\Big[1 - D_{\mathbf{M}}^{\boldsymbol{\phi}}\Big(\mathbf{x}_{b_i}, (1-t_{b_i})m_{b_i} + t_{b_i} G_{\mathbf{M}}^{\boldsymbol{\theta},(0)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}), t_{b_i} m_{b_i} + (1-t_{b_i}) G_{\mathbf{M}}^{\boldsymbol{\theta},(1)}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i})\Big)\Big] \\
&\quad + \alpha_1 \Big| G_{\mathbf{M}}^{\boldsymbol{\theta},(t_{b_i})}(\mathbf{z}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}) - m_{b_i} \Big|^2 \Big\} \Bigg\}, \\
&\bigtriangledown_{\boldsymbol{\zeta}} \Bigg\{ \frac{1}{B_1}\sum_{i=1}^{B_1} \Big\{ t_{b_i} \log D_{\mathbf{Y}}^{\boldsymbol{\xi}}\Big(\mathbf{x}_{b_i}, m_{b_i}, (1-t_{b_i})y_{b_i} + t_{b_i} G_{\mathbf{Y}}^{\boldsymbol{\zeta},(0)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}), t_{b_i} y_{b_i} + (1-t_{b_i}) G_{\mathbf{Y}}^{\boldsymbol{\zeta},(1)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i})\Big) \\
&\quad + (1-t_{b_i}) \log\Big[1 - D_{\mathbf{Y}}^{\boldsymbol{\xi}}\Big(\mathbf{x}_{b_i}, m_{b_i}, (1-t_{b_i})y_{b_i} + t_{b_i} G_{\mathbf{Y}}^{\boldsymbol{\zeta},(0)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}), t_{b_i} y_{b_i} + (1-t_{b_i}) G_{\mathbf{Y}}^{\boldsymbol{\zeta},(1)}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i})\Big)\Big] \Big\} \\
&\quad + \alpha_4 \frac{1}{B_1}\sum_{i=1}^{B_1} \Big| G_{\mathbf{Y}}^{\boldsymbol{\zeta},(t_{b_i})}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, y_{b_i}) - y_{b_i} \Big| + \alpha_4 \frac{1}{B_2} \sum_{i=B_1+1}^{B} \max\Big\{0, c_{b_i} - G_{\mathbf{Y}}^{\boldsymbol{\zeta},(t_{b_i})}(\widetilde{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, t_{b_i}, m_{b_i}, c_{b_i})\Big\}
\end{aligned}$$

end while

Second Step (after $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$ and $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$ have been fully trained):

while not converged **do**

- Compute $\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T=1, m_i), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T=0, m_i), \widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=1, m_i, \widetilde{y}_i), \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=0, m_i, \widetilde{y}_i), \mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}(\widehat{\mathbf{z}}_i, \mathbf{x}_i) = (I_{\mathbf{M}}^{\boldsymbol{\psi},(0)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i), I_{\mathbf{M}}^{\boldsymbol{\psi},(1)}(\widehat{\mathbf{z}}_i, \mathbf{x}_i))$, and $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i) = (I_{\mathbf{Y}}^{\boldsymbol{\varphi},(0)}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i), I_{\mathbf{Y}}^{\boldsymbol{\varphi},(1)}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i))$. Let $(\overline{m}_i^{(0)}, \overline{m}_i^{(1)}) = t_i \cdot (\widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T=1, m_i), m_i) + (1-t_i) \cdot (m_i, \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T=0, m_i))$, $(\overline{y}_i^{(0)}, \overline{y}_i^{(1)}) = t_i \cdot (\widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=1, m_i, y_i), y_i) + (1-t_i) \cdot (y_i, \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=0, m_i, y_i))$, and $S_{It} = \{\mathbf{x}_i, t_i, m_i, \widetilde{y}_i, \delta_i, \widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T=1, m_i), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T=0, m_i), \widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=1, m_i, \widetilde{y}_i), \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=0, m_i, \widetilde{y}_i), \mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}(\widehat{\mathbf{z}}_i, \mathbf{x}_i), \mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i), i=1,\ldots,n\} = S_{It1} \cup S_{It2}$, where $S_{It1} = \{\mathbf{x}_i, t_i, m_i, y_i, \delta_i = 1, \widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T=1, m_i), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T=0, m_i), \widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=1, m_i, y_i), \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=0, m_i, y_i), \mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}(\widehat{\mathbf{z}}_i, \mathbf{x}_i),$

$\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i), i = 1, \ldots, n_1\}$, and $S_{It2} = \{\mathbf{x}_i, t_i, m_i, c_i, \delta_i = 0, \widehat{G}_{\mathbf{M}}^{(0)}(\mathbf{z}_i, \mathbf{x}_i, T = 1, m_i), \widehat{G}_{\mathbf{M}}^{(1)}(\mathbf{z}_i, \mathbf{x}_i, T = 0, m_i), \widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T = 1, m_i, c_i), \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T = 0, m_i, c_i), \mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}(\widehat{\mathbf{z}}_i, \mathbf{x}_i), \mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i), i = n_1 + 1, \ldots, n\}$.

- Randomly select B samples from S_{It}, where B_1 samples from S_{It1}, B_2 samples from S_{It2}, and $B = B_1 + B_2$. Denote the subscripts of the selected samples by $\{b_i : i = 1, \ldots, B\}$, $\{b_i : i = 1, \ldots, B_1\}$, and $\{b_i : i = B_1 + 1, \ldots, B\}$.

- Update $D_{\mathbf{I_M}}^{\boldsymbol{\omega}}$ and $D_{\mathbf{I_Y}}^{\boldsymbol{\lambda}}$ by ascending their stochastic gradients:

$$\nabla_{\boldsymbol{\omega}} \left\{ \frac{1}{B} \sum_{i=1}^{B} \left\{ \log D_{\mathbf{I_M}}^{\boldsymbol{\omega}} \left(\mathbf{x}_{b_i}, \overline{m}_{b_i}^{(0)}, \overline{m}_{b_i}^{(1)} \right) + \log \left[1 - D_{\mathbf{I_M}}^{\boldsymbol{\omega}} \left(\mathbf{x}_{b_i}, I_{\mathbf{M}}^{\boldsymbol{\psi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}), I_{\mathbf{M}}^{\boldsymbol{\psi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}) \right) \right] \right\} \right\}$$

$$\nabla_{\boldsymbol{\lambda}} \left\{ \frac{1}{B_1} \sum_{i=1}^{B_1} \left\{ \log D_{\mathbf{I_Y}}^{\boldsymbol{\lambda}} \left(\mathbf{x}_{b_i}, m_{b_i}, \overline{y}_{b_i}^{(0)}, \overline{y}_{b_i}^{(1)} \right) + \log \left[1 - D_{\mathbf{I_Y}}^{\boldsymbol{\lambda}} \left(\mathbf{x}_{b_i}, m_{b_i}, I_{\mathbf{Y}}^{\boldsymbol{\varphi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}), I_{\mathbf{Y}}^{\boldsymbol{\varphi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) \right) \right] \right\} \right\}$$

- Update $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}$ and $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$ by descending its stochastic gradient:

$$\begin{aligned}
&\nabla_{\boldsymbol{\psi}} \left\{ \frac{1}{B} \sum_{i=1}^{B} \left\{ \log D_{\mathbf{I_M}}^{\boldsymbol{\omega}} \left(\mathbf{x}_{b_i}, \overline{m}_{b_i}^{(0)}, \overline{m}_{b_i}^{(1)} \right) + \log \left[1 - D_{\mathbf{I_M}}^{\boldsymbol{\omega}} \left(\mathbf{x}_{b_i}, I_{\mathbf{M}}^{\boldsymbol{\psi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}), I_{\mathbf{M}}^{\boldsymbol{\psi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}) \right) \right] \right. \right. \\
&\quad + \alpha_2 \left| \left(\overline{m}_{b_i}^{(0)} - \overline{m}_{b_i}^{(1)} \right) - \left(I_{\mathbf{M}}^{\boldsymbol{\psi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}) - I_{\mathbf{M}}^{\boldsymbol{\psi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}) \right) \right| + \alpha_3 \left| I_{\mathbf{M}}^{\boldsymbol{\psi},(t_{b_i})}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}) - m_{b_i} \right|^2 \left. \right\} \Bigg\} \\
&\nabla_{\boldsymbol{\varphi}} \left\{ \frac{1}{B_1} \sum_{i=1}^{B_1} \left\{ \log D_{\mathbf{I_Y}}^{\boldsymbol{\lambda}} \left(\mathbf{x}_{b_i}, m_{b_i}, \overline{y}_{b_i}^{(0)}, \overline{y}_{b_i}^{(1)} \right) + \log \left[1 - D_{\mathbf{I_Y}}^{\boldsymbol{\lambda}} \left(\mathbf{x}_{b_i}, m_{b_i}, I_{\mathbf{Y}}^{\boldsymbol{\varphi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}), I_{\mathbf{Y}}^{\boldsymbol{\varphi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) \right) \right] \right\} \right. \\
&\quad + \alpha_5 \frac{1}{B_1} \sum_{i=1}^{B_1} \left| \left(\overline{y}_{b_i}^{(0)} - \overline{y}_{b_i}^{(1)} \right) - \left(I_{\mathbf{Y}}^{\boldsymbol{\varphi},(0)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) - I_{\mathbf{Y}}^{\boldsymbol{\varphi},(1)}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) \right) \right| \\
&\quad + \alpha_6 \frac{1}{B_1} \sum_{i=1}^{B_1} \left| I_{\mathbf{Y}}^{\boldsymbol{\varphi},(t_{b_i})}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) - y_{b_i} \right| + \alpha_6 \frac{1}{B_2} \sum_{i=B_1+1}^{B} \max \left\{ 0, c_{b_i} - I_{\mathbf{Y}}^{\boldsymbol{\varphi},(t_{b_i})}(\overline{\mathbf{z}}_{b_i}, \mathbf{x}_{b_i}, m_{b_i}) \right\} \Bigg\}
\end{aligned}$$

end while

S6 Competing Methods

This section briefly introduces five competing methods: LR+AFT, LR+2AFT, ILR+IAFT, RF+RSF, and BART.

S6.1 Linear regression + AFT interaction model (LR+AFT)

As our proposed method consists of a mediator layer and an outcome layer, the competing approaches should also contain the two components.

LR+AFT adopts linear regression (LR) in the mediator layer and an AFT interaction model in the outcome layer. The LR is implemented using the *LinearRegression* function available in the *sklearn.linear_model* module of Python. We now introduce the AFT interaction model (Lo, 2002; Tabib and Larocque, 2020). It is a generalization of the methodology presented in Tabib and Larocque (2020), which builds upon the work of Lo (2002). Lo (2002) employed a logistic regression model to estimate individualized treatment effects without mediators, focusing on binary responses. They included all covariates, the binary treatment variable, and interactions between the treatment variable and covariates. Tabib and Larocque (2020) adapted this method using an AFT model instead of logistic regression. We generalize their idea to incorporate mediators into the AFT model as

follows:

$$\log Y(\mathbf{x}, t, M_t(\mathbf{x})) = \beta_0 + \mathbf{x}'\boldsymbol{\beta}_1 + t\beta_2 + M_t(\mathbf{x})\beta_3 + \mathbf{x}'t\boldsymbol{\beta}_4 + \epsilon. \quad \text{(S6.53)}$$

We use the function *survreg* in the *survival* package (Therneau et al., 2015) of R and fit models with the exponential and lognormal distributions in this study.

Once the model is fitted, we can follow a three-step process to estimate ICEs. First, we predict the potential mediators by performing LR while setting the treatment variable to zero or one. This step allows us to obtain estimates of the potential mediator values. Second, we use these predicted potential mediators to forecast survival times based on an AFT interaction model. This step helps us generate predictions for the potential event times. Finally, we use these predicted potential event times to estimate ICEs.

S6.2 Linear regression + 2AFT interaction model (LR+2AFT)

LR+2AFT also incorporates LR in the mediator layer, utilizing the *LinearRegression* function from the *sklearn.linear_model* module in Python. However, it generalizes the approach of Tabib and Larocque (2020) by fitting two separate AFT models (2AFT) as follows:

$$\log Y(\mathbf{x}, 0, M_t(\mathbf{x})) = \beta_0 + \mathbf{x}'\boldsymbol{\beta}_1 + M_t(\mathbf{x})\beta_2 + \epsilon, \quad \text{(S6.54)}$$

$$\log Y(\mathbf{x}, 1, M_t(\mathbf{x})) = \beta_3 + \mathbf{x}'\boldsymbol{\beta}_4 + M_t(\mathbf{x})\beta_5 + \epsilon. \quad \text{(S6.55)}$$

Similarly, we use the function *survreg* in the *survival* package (Therneau et al., 2015) of R and fit models with the exponential and lognormal distributions. Once the model is fitted, we can first predict the potential mediators through LR, then use these potential mediators to predict the potential event times through the 2AFT model, and finally estimate ICEs.

S6.3 Interaction Linear regression + Another Interaction AFT model (ILR+IAFT)

ILR+IAFT incorporates interaction LR in the mediator layer using the *LinearRegression* function from the *sklearn.linear_model* module in Python. The formula for the mediator layer is as follows:

$$M_t(\mathbf{x}) = \beta_{M0} + \mathbf{x}'\boldsymbol{\beta}_{M1} + t\beta_{M2} + \mathbf{x}'t\boldsymbol{\beta}_{M3} + \epsilon. \tag{S6.56}$$

In the outcome layer, we use another interaction AFT model, which can be fitted using the *survreg* function in the *survival* package (Therneau et al., 2015) of R. The formula for the outcome layer is as follows:

$$\log Y(\mathbf{x}, t, M_t(\mathbf{x})) = \beta_0 + \mathbf{x}'\boldsymbol{\beta}_1 + t\beta_2 + M_t(\mathbf{x})\beta_3 + \mathbf{x}'t\boldsymbol{\beta}_4 + \mathbf{x}'M_t(\mathbf{x})\boldsymbol{\beta}_5 + tM_t(\mathbf{x})\beta_6 + \epsilon. \tag{S6.57}$$

Subsequently, once the model is fitted, we first fit the ILR model to predict the potential mediators and then use the predicted mediators to fit the IAFT model to predict potential event times. Finally, we can estimate ICEs.

S6.4 Random forest (RF) + Random survival forest (RSF) (RF+RSF)

This method uses RF for regression in the mediator layer and RSF in the outcome layer. We use the *RandomForestRegressor* function provided by the *sklearn.ensemble* module in Python in the mediator layer. For the outcome layer, the function *survival_forest* in the R package *grf* https://github.com/grf-labs/grf is used. Once the model is fitted, an alternative approach is to predict the potential mediators through RF for regression. However, it is worth noting that the *survival_forest* function does not directly predict potential event times for new data. Instead, it provides the conditional survival function. Nonetheless, we can still estimate ICEs using the following methods. Denote by $S^t(\cdot, \mathbf{x}, M_{t'}(\mathbf{x}))$ the conditional survival function of the event time for a subject, for each $t, t' \in \{0, 1\}$, $S^t(y, \mathbf{x}, M_{t'}(\mathbf{x})) = P(Y > y|t, \mathbf{x}, M_{t'}(\mathbf{x}))$, then we have

$$\begin{aligned}
\xi(t;\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x}, t, M_1(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, t, M_0(\mathbf{x}))] = \int_0^\infty [S^t(y, \mathbf{x}, M_1(\mathbf{x})) - S^t(y, \mathbf{x}, M_0(\mathbf{x}))]dy, \\
\zeta(t;\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x}, 1, M_t(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, 0, M_t(\mathbf{x}))] = \int_0^\infty [S^1(y, \mathbf{x}, M_t(\mathbf{x})) - S^0(y, \mathbf{x}, M_t(\mathbf{x}))]dt, \\
\tau(\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x}, 1, M_1(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, 0, M_0(\mathbf{x}))] = \int_0^\infty [S^1(y, \mathbf{x}, M_1(\mathbf{x})) - S^0(y, \mathbf{x}, M_0(\mathbf{x}))]dt.
\end{aligned} \tag{S6.58}$$

Thus, we can estimate ICEs by approximating a definite integral of the predicted conditional survival function.

S6.5 Bayesian additive regression trees (BART)

BART is described in Sparapani et al. (2021), which is implemented using the functions *wbart* and *abart* in the R package *BART* in the mediator and

outcome layers, respectively.

Once the model is fitted, an alternative approach is to predict the potential mediators using the *wbart* function. However, it is important to note that the *abart* function does not directly predict potential event times for new data. Instead, it provides the logarithm of the predicted potential event times. We need to apply the exponential function to the logarithmic predictions to obtain the predicted potential event times. By doing so, we can proceed to estimate ICEs.

S7 Performance Metrics

In the absence of mediators, if both factual and counterfactual outcomes are observed, but the underlying distribution is unknown, Yoon et al. (2018) introduced an empirical precision in the estimation of heterogeneous effect (PEHE) as follows:

$$\hat{\epsilon}_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}([y_i(1) - y_i(0)] - [\hat{y}_i(1) - \hat{y}_i(0)])^2, \tag{S7.59}$$

where $y_i(1)$ and $y_i(0)$ are potential outcomes of treated and controlled, respectively, and $\hat{y}_i(1)$ and $\hat{y}_i(0)$ are their estimates. We generalize (S7.59) to define three metrics about the ICEs defined in Section 2.3 in the paper with potential survival time.

We note the observed dataset is $S = \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = \widetilde{y}_i, \delta = \delta_i\}_{i=1}^{n}$, where n is the number of observations. Let $S_r = \{\mathbf{x}_{ri}, t_{ri}, m_{ri}, \widetilde{y}_{ri}, \delta_{ri}\}_{i=1}^{n_r}$ denote the observed training dataset and $S_e = \{\mathbf{x}_{ei}, t_{ei}, m_{ei}, \widetilde{y}_{ei}, \delta_{ei}\}_{i=1}^{n_e}$ denote the observed testing dataset, where n_r and n_e are the numbers of training and testing samples, respectively, with $n_r + n_e = n$.

For prediction, we first generate $\widehat{n}_e$ samples $\{\widehat{\mathbf{z}}_{eh}, h = 1, \ldots, \widehat{n}_e\}$ from $\widehat{\mathbf{Z}} \sim P_{\widehat{\mathbf{Z}}}$ and $\overline{n}_e$ samples $\{\overline{\mathbf{z}}_{ej}, j = 1, \ldots, \overline{n}_e\}$ from $\overline{\mathbf{Z}} \sim P_{\overline{\mathbf{Z}}}$, and calculate conditional samples $\{\widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh}, \mathbf{x}_{ei}), h = 1, \ldots, \widehat{n}_e\}$ and $\{\widehat{I}_{\mathbf{Y}}^{(\bar{t})}(\overline{\mathbf{z}}_{ej}, \mathbf{x}_{ei}, \widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh}, \mathbf{x}_{ei})), h = 1, \ldots, \widehat{n}_e\, j = 1, \ldots, \overline{n}_e\}$ for each $t, \bar{t} \in \{0, 1\}$. We take $\widehat{n}_e = \overline{n}_e = 100$. Then, we can estimate $\xi(t; \mathbf{x}_{ei}), \zeta(t; \mathbf{x}_{ei})$, and $\tau(\mathbf{x}_{ei})$ based on (2.7), (2.8), and (2.9) in the paper, denote as $\widehat{\xi}(t; \mathbf{x}_{ei}), \widehat{\zeta}(t; \mathbf{x}_{ei})$, and $\widehat{\tau}(\mathbf{x}_{ei})$. Then, the metrics on the testing dataset are defined by

$$\begin{aligned}
\hat{\epsilon}_{\text{PEHE}_{\text{TE}}} &= \frac{1}{n_e} \sum_{i=1}^{n_e} \Big\{ \tau(\mathbf{x}_{ei}) - \widehat{\tau}(\mathbf{x}_{ei}) \Big\}^2, \\
\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}} &= \frac{1}{n_e} \sum_{i=1}^{n_e} \Big\{ \xi(t; \mathbf{x}_{ei}) - \widehat{\xi}(t; \mathbf{x}_{ei}) \Big\}^2, \\
\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}} &= \frac{1}{n_e} \sum_{i=1}^{n_e} \Big\{ \zeta(t; \mathbf{x}_{ei}) - \widehat{\zeta}(t; \mathbf{x}_{ei}) \Big\}^2.
\end{aligned}$$

For simplicity, we consider the decomposition with $\xi(1; \mathbf{x}_{ei})$ and $\zeta(0; \mathbf{x}_{ei})$, but the alternative decomposition with $\xi(0; \mathbf{x}_{ei})$ and $\zeta(1; \mathbf{x}_{ei})$ can also be used with similar procedures. A small value of $\hat{\epsilon}_{\text{PEHE}}$ means an accurate estimate.

S8 Hyperparameters of CGAN-ICMA-SO

Table S1 presents the setting of the hyperparameters in the network for the simulation studies and the application.

Table S1: Hyperparameters of CGAN-ICMA-SO

Blocks	Setting of the hyper-parameters in the network
Initialization	Weight matrix: Xavier Initialization. Bias vector: Zero initialization.
Batch size (B)	256
Depth of layers	3
Hidden state dimension (all blocks)	10
$\alpha_1, \ldots, \alpha_6$	1
Optimization	Adam Moment Optimization

S9 Simulation Implementation and Results

We generate 1,000 samples and use 900 instances for training and 100 cases for testing (the training rate is 0.9). We compare our approach with the five competing methods by repeating them 100 times and reporting the average value of the square root of metrics and the corresponding standard deviation (std). Table S2 presents comparison results when CR = 50%, with the last column indicating the time spent in each replication. Our method per-

forms best with the smallest values for the averaged $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$, $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$, and $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$, suggesting that our approach estimates ICEs with survival outcomes more accurately than the others, although it takes longer.

Table S2: Performance of five methods for estimating ICEs (CR = 50%)

Methods	Mean(std) based on 100 replications			
	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$	time(s)
CGAN-ICMA-SO	6.519(2.543)	2.134(1.192)	6.040(2.559)	107.87
LR+AFT(exponential)	97.943(47.070)	72.037(36.917)	27.357(11.279)	0.78
LR+AFT(lognormal)	9.490(2.920)	3.934(1.020)	8.449(3.099)	0.66
LR+2AFT(exponential)	99.917(49.080)	72.562(37.588)	28.779(12.763)	0.35
LR+2AFT(lognormal)	9.509(3.054)	3.839(0.992)	8.437(3.151)	0.39
ILR+IAFT(exponential)	390.969(651.824)	131.966(193.621)	297.544(482.996)	0.63
ILR+IAFT(lognormal)	7.301(2.760)	2.927(0.755)	6.428(2.748)	0.52
RF+RSF	11.249(3.004)	5.605(0.495)	8.824(3.300)	1.71
BART	126.938(69.397)	102.821(55.811)	52.084(33.726)	8.14

Note: The reported time includes training and prediction. CPU time was used for CGAN-ICMA-SO, and GPU parallelization can potentially reduce the time.

Notably, our goal is to compare our proposed method with existing approaches to evaluate its empirical performance. However, since no methods are specifically designed to address the problem at hand, we conducted simulations and proposed the above alternative approaches for comparison. It is crucial to highlight that the alternative methods used for comparison fail to achieve identification of the ICEs. This limitation stems from the fact that in our method, as demonstrated by Equation (2.4) in the paper,

we employ a sampling-based approach by drawing multiple values from the estimated probability distribution of the potential mediator, $P_{M(\mathbf{x},t)}$, to estimate potential outcomes. On the contrary, the competing methods rely on separate regressions for the mediator and outcome layers, incorporating the predicted expected value of the mediator into the outcome regression model for prediction, failing to satisfy the requirements for identification. By adopting the sampling-based approach, our method provides a comprehensive understanding of the causal mediation effects. It allows for a more nuanced analysis, capturing the potential variations and uncertainties in the estimation process. Despite the longer computational time required, we find the trade-off acceptable due to the substantial improvement in accuracy. Moreover, the increased accuracy of our model contributes to a more reliable assessment of the causal mediation effects, which is paramount in understanding the underlying mechanisms and making informed decisions.

We then examine the robustness of the proposed and competing methods to model parameters using the setting in the paper. Tables S3 and S4 show the performance of CGAN-ICMA-SO when changing CR to 30%, but other settings remain the same and changing the batch size from 256 to 128, $\alpha_1, \ldots, \alpha_6$ from 1 to 1.5, layer depth from 3 to 2, and the hidden state dimension from 10 to 8 but CR remains 50%. The last column reports the

time spent in each replication. CGAN-ICMA-SO still outperforms the five other methods in almost all situations.

Table S3: Performance of six methods for estimating ICEs (CR = 30%)

Methods	Mean(std) based on 100 replications			
	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$	time(s)
CGAN-ICMA-SO	6.201(2.676)	1.982(1.338)	5.748(2.592)	107.91
LR+AFT(exponential)	15.562(3.705)	10.230(3.190)	8.994(2.722)	0.49
LR+AFT(lognormal)	8.733(3.136)	2.766(0.684)	8.448(3.182)	0.49
LR+2AFT(exponential)	15.626(3.785)	10.253(3.197)	9.029(2.742)	0.58
LR+2AFT(lognormal)	8.866(3.187)	2.782(0.712)	8.455(3.199)	0.87
ILR+IAFT(exponential)	18.894(6.763)	9.256(2.552)	13.733(5.908)	0.70
ILR+IAFT(lognormal)	7.054(2.799)	2.422(0.625)	6.441(2.797)	0.55
RF+RSF	11.194(2.791)	6.930(0.832)	8.695(3.295)	2.15
BART	11.124(3.068)	8.422(1.807)	6.031(2.715)	12.23

We also increase the sample size n from 1000 to 2000 while keeping the other settings unchanged and decrease the training rate from 0.9 to 0.8 while maintaining the other parameters the same. Table S5 shows the results. Again, CGAN-ICMA-SO is superior to others regardless of the situations considered. Moreover, the performance of most methods improves when the sample size or training rate increases.

Next, we consider another setting to demonstrate our method's superi-

Table S4: Performance of six methods for estimating ICEs (CR = 50%)

Methods	Mean(std) based on 100 replications			
	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$	time(s)
CGAN-ICMA-SO (Table 1)	6.519(2.543)	2.134(1.192)	6.040(2.559)	107.87
CGAN-ICMA-SO (batch size=128)	6.939(2.923)	2.358(1.189)	6.316(2.951)	102.14
CGAN-ICMA-SO ($\alpha_1, \ldots, \alpha_6$=1.5)	6.785(2.372)	2.550(1.317)	6.090(2.487)	106.51
CGAN-ICMA-SO (layer depth 2)	6.878(2.786)	2.837(1.160)	6.043(2.853)	100.69
CGAN-ICMA-SO (hsd = 8)	6.918(2.549)	2.867(1.264)	5.997(2.703)	102.95
LR+AFT(exponential)	97.943(47.070)	72.037(36.917)	27.357(11.279)	0.78
LR+AFT(lognormal)	9.490(2.920)	3.934(1.020)	8.449(3.099)	0.66
LR+2AFT(exponential)	99.917(49.080)	72.562(37.588)	28.779(12.763)	0.35
LR+2AFT(lognormal)	9.509(3.054)	3.839(0.992)	8.437(3.151)	0.39
ILR+IAFT(exponential)	390.969(651.824)	131.966(193.621)	297.544(482.996)	0.63
ILR+IAFT(lognormal)	7.301(2.760)	2.927(0.755)	6.428(2.748)	0.52
RF+RSF	11.249(3.004)	5.605(0.495)	8.824(3.300)	1.71
BART	126.938(69.397)	102.821(55.811)	52.084(33.726)	8.14

Note: "hsd" — hidden state dimension.

ority to a greater extent. The model is defined as follows:

$$M(\mathbf{x}, t) = 0.2 + 0.5|x_2| + 1.5x_5^2 + 0.1x_6 + t(x_3 + x_4)^2 + \epsilon_1,$$

$$Y(\mathbf{x}, t, m_t(\mathbf{x})) = 0.1 + 0.2\exp(x_{10}) + 2|x_5| + t(x_8 + x_9)^2 + 0.5m_t^2(\mathbf{x}) + \epsilon_2,$$

$$\widetilde{Y}(\mathbf{x}, t, m_t(\mathbf{x})) = \min(Y(\mathbf{x}, t, m_t(\mathbf{x})), C),\ \delta = \mathbb{I}\{Y(\mathbf{x}, t, m_t(\mathbf{x})) < C\},$$

where the notation and distribution setup is the same as in the paper.

Again, we generate 1,000 samples and use 900 instances for training and 100 cases for testing. Table S6 presents the comparison results when CR =

Table S5: Performance of six methods for estimating ICEs (CR = 50%)

Methods	Mean(std) based on 100 replications		
	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$
$n = 2000$, training rate = 0.9			
CGAN-ICMA-SO	5.419(2.200)	1.782(1.103)	4.916(2.148)
LR+AFT(exponential)	93.117(31.533)	67.869(24.160)	26.631(7.925)
LR+AFT(lognormal)	9.214(2.078)	3.949(0.795)	8.165(2.361)
LR+2AFT(exponential)	95.214(31.060)	68.363(23.885)	28.192(8.173)
LR+2AFT(lognormal)	9.235(2.212)	3.862(0.794)	8.149(2.410)
ILR+IAFT(exponential)	631.037(2725.948)	139.560(273.889)	524.544(2452.434)
ILR+IAFT(lognormal)	7.124(2.077)	3.005(0.556)	6.256(2.177)
RF+RSF	10.690(2.283)	5.503(0.367)	8.377(2.554)
BART	320.647(230.179)	238.454(182.446)	165.136(117.392)
$n = 1000$, training rate = 0.8			
CGAN-ICMA-SO	7.076(2.217)	2.363(1.428)	6.561(2.131)
LR+AFT(exponential)	102.948(43.433)	75.581(32.944)	28.809(11.972)
LR+AFT(lognormal)	9.583(2.048)	4.069(0.849)	8.528(2.270)
LR+2AFT(exponential)	105.300(45.811)	76.274(33.821)	30.435(13.533)
LR+2AFT(lognormal)	9.609(2.124)	3.972(0.828)	8.517(2.312)
ILR+IAFT(exponential)	588.769(1293.877)	165.603(242.278)	503.405(1215.426)
ILR+IAFT(lognormal)	7.437(2.009)	3.007(0.618)	6.571(2.079)
RF+RSF	11.328(2.352)	5.723(0.359)	8.908(2.462)
BART	109.825(52.659)	90.640(43.212)	43.289(24.245)

50%. Our method significantly outperforms others regarding the averaged $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$, $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$, and $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$, suggesting that our approach estimates all three metrics with survival outcomes more accurately than the

others.

Table S6: Performance of six methods for estimating ICEs in other setting (CR = 50%)

Methods	Mean(std) based on 100 replications			
	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$	$\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$	time(s)
CGAN-ICMA-SO	5.122(1.765)	2.691(1.217)	4.232(1.493)	149.88
LR+AFT(exponential)	428.666(373.617)	244.553(224.063)	191.517(152.674)	0.74
LR+AFT(lognormal)	15.045(5.788)	12.181(5.900)	7.167(1.532)	0.83
LR+2AFT(exponential)	445.754(413.446)	253.348(241.264)	200.369(174.837)	0.70
LR+2AFT(lognormal)	15.057(5.973)	12.672(6.557)	6.844(1.282)	0.93
ILR+IAFT(exponential)	$9.351\times10^3(2.034\times10^4)$	$2.686\times10^3(6.315\times10^3)$	$7.315\times10^3(1.578\times10^4)$	0.60
LR+IAFT(lognormal)	10.679(3.121)	8.588(2.828)	4.570(0.945)	0.57
RF+RSF	10.236(1.408)	6.153(0.552)	6.316(1.387)	2.25
BART	$1.266\times10^3(1.490\times10^3)$	$1.097\times10^3(1.253\times10^3)$	349.601(472.022)	11.47

S10 Additional Results in ADNI Study

S10.1 Results using five other methods

We use the five other methods to estimate $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs defined in the article. The results shown in the following figures are discouraging. We noticed that the other five methods all performed poorly. For the predicted values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$, the values estimated by LR+AFT (exponential), LR+AFT (lognormal), LR+2AFT (exponential), and LR+2AFT (lognormal) are completely invariant in each fold, and the values estimated by ILR+IAFT (exponential), ILR+IAFT (lognormal), RF+RSF, and BART randomly fluctuated from positive to negative with some zero values. For the predicted values of three ICEs, the five methods all produce a significant amount of positive values which are unreasonable, and LR+AFT (exponential), LR+AFT (lognormal), LR+2AFT (exponential), LR+2AFT (lognormal), ILR+IAFT (exponential), and ILR+IAFT (exponential) produce some outliers.

LR+AFT (exponential):

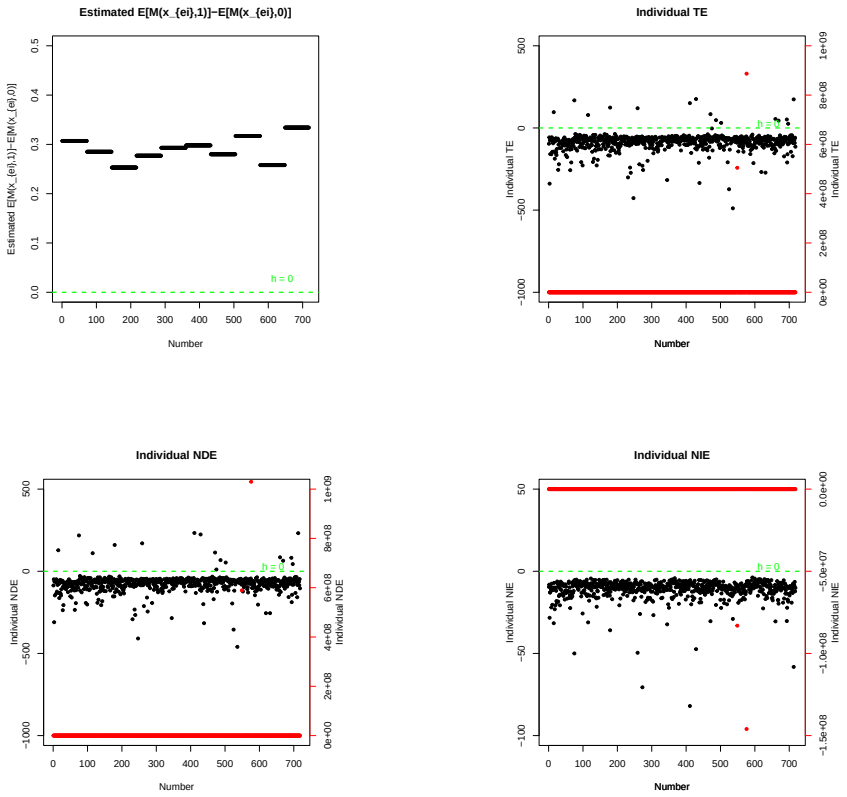

Figure S1: The estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

LR+AFT (lognormal):

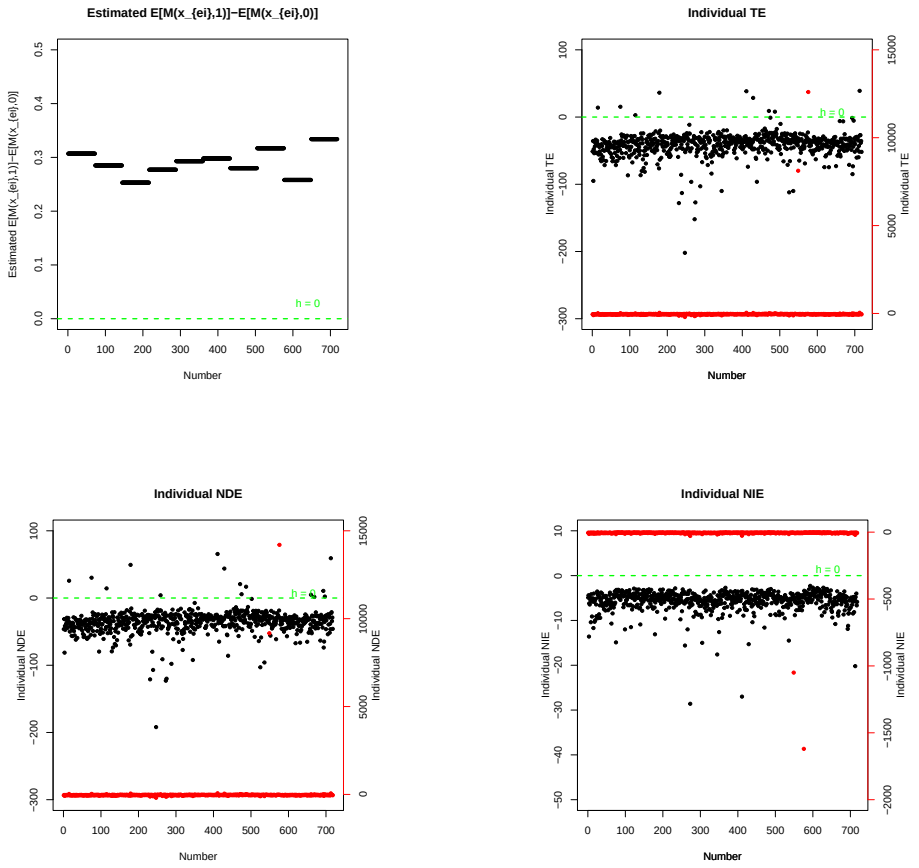

Figure S2: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

LR+2AFT (exponential):

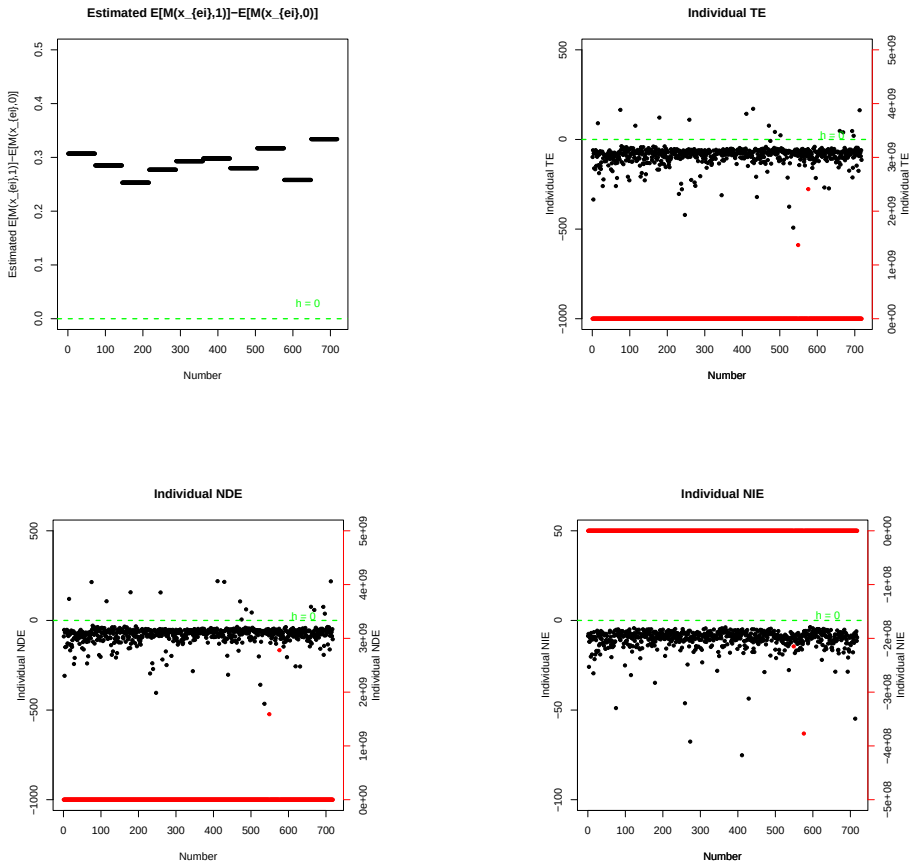

Figure S3: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

LR+2AFT (lognormal):

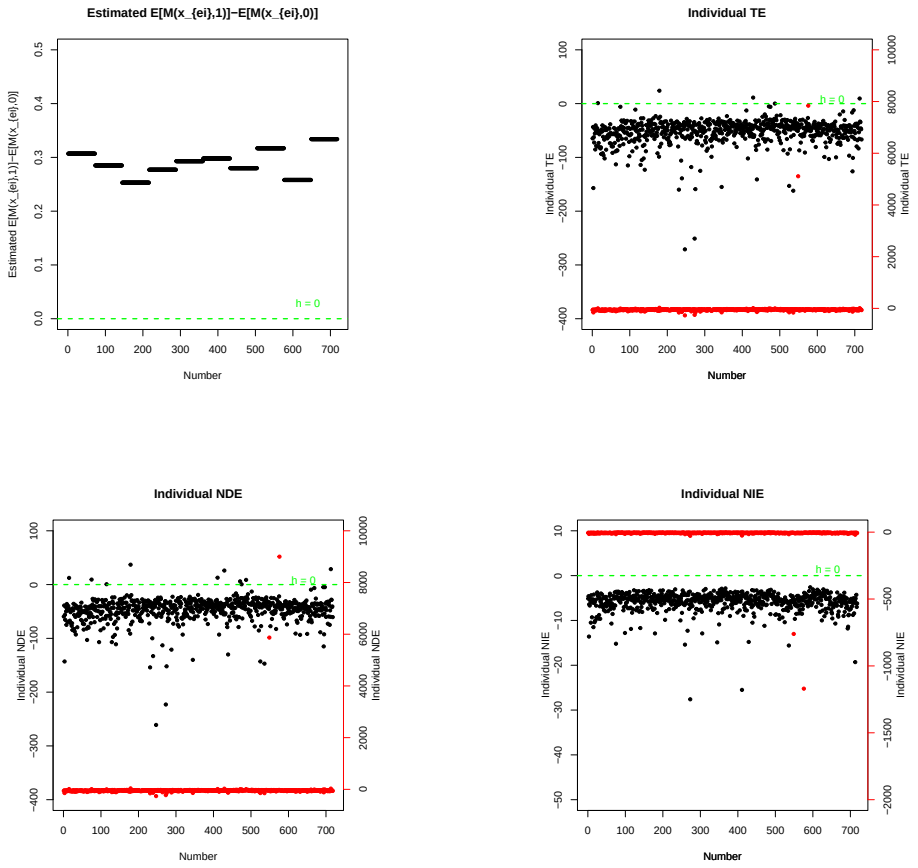

Figure S4: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

ILR+IAFT (exponential):

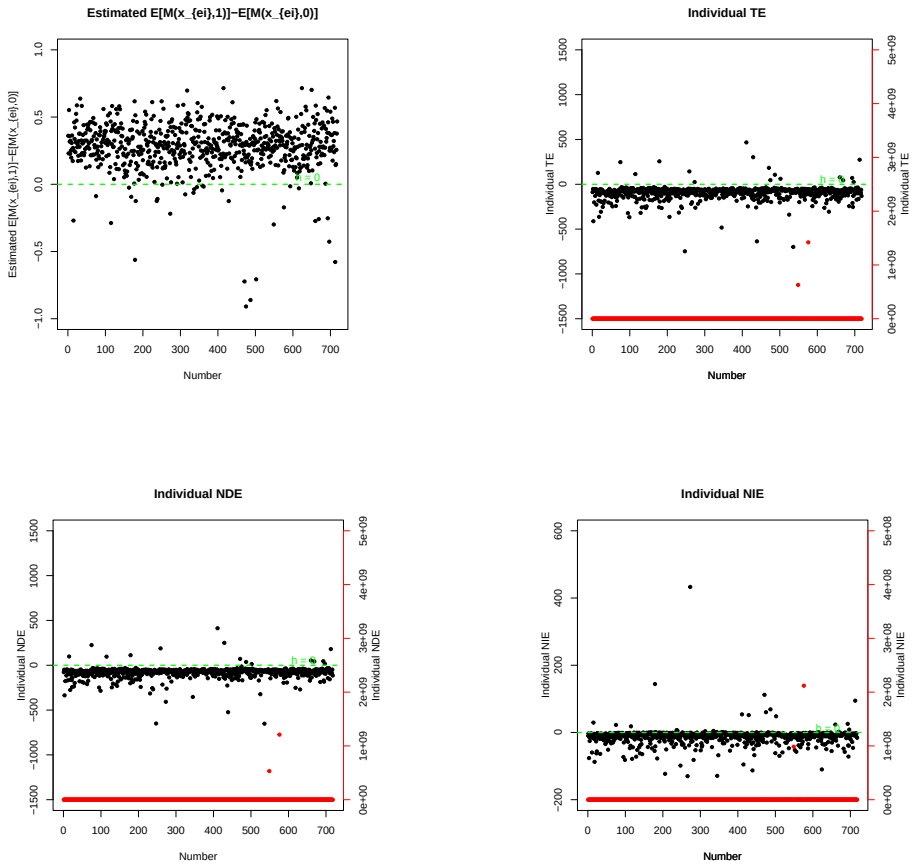

Figure S5: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

ILR+IAFT (lognormal):

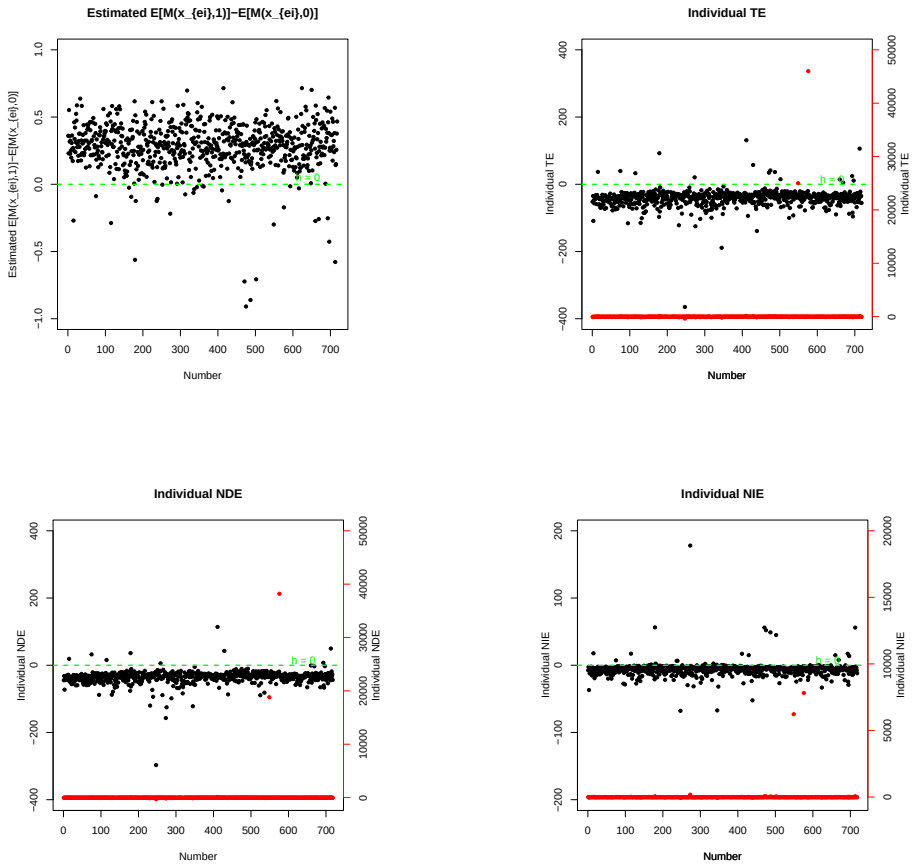

Figure S6: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

RF+RSF:

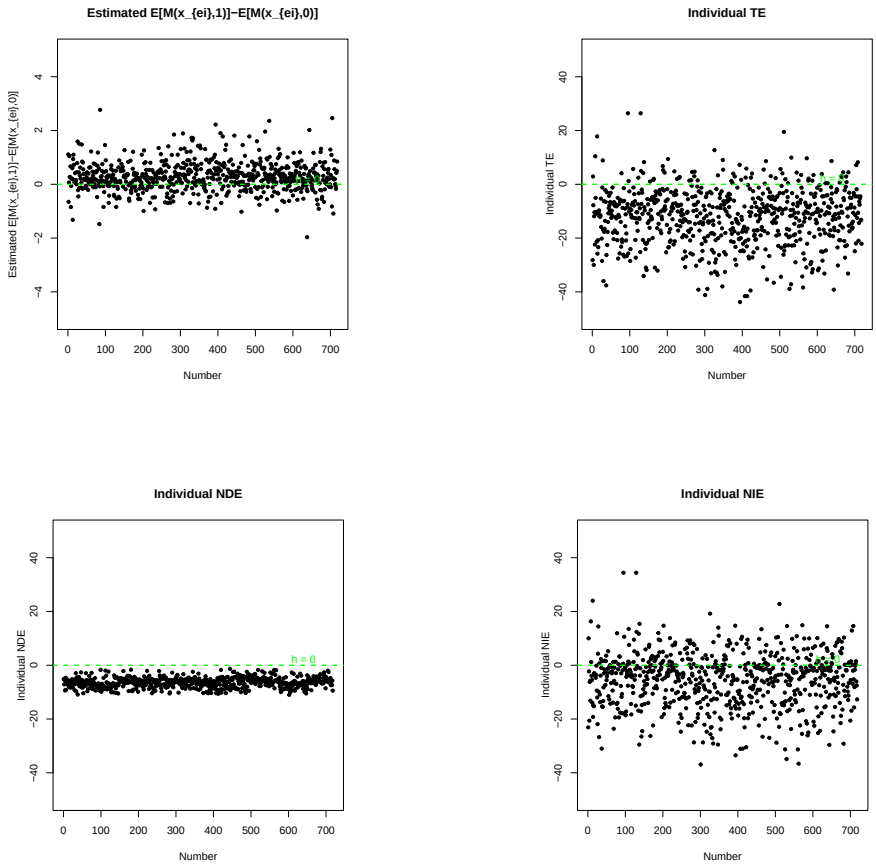

Figure S7: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

BART:

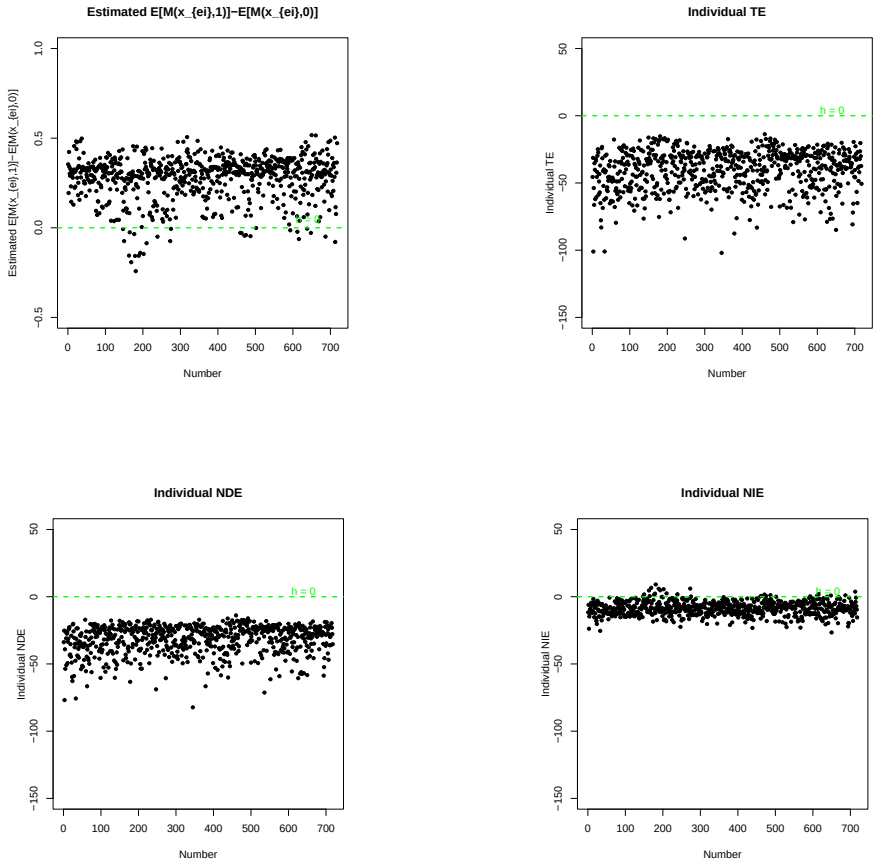

Figure S8: Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and three ICEs with respect to the patient index.

S10.2 Best linear prediction of GACEs

We model GACEs using the multivariate OLS regression with six covariates. Although it may be misspecified, the model gives the best linear predictor of GACEs with six covariates and provides an accessible summary of the effect heterogeneities. Table S7 shows the results, which agree with those obtained in the paper. For example, in the white group, the group average TE is significant and negative, suggesting that the APOE-ε4–AD association is stronger in the white group. Similarly, the coefficients of gender, Hispanic or Latino, married, education level, and age yield the same conclusions as above.

S10.3 Average causal effects

We can obtain the average causal effects based on the three kinds of estimated ICEs. The results are: average TE $= \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}(\mathbf{x}_{ei}) = -8.332$, average NDE $= \frac{1}{n}\sum_{i=1}^{n}\widehat{\zeta}(0;\mathbf{x}_{ei}) = -3.205$, and average NIE $= \frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}(1;\mathbf{x}_{ei})$ $= -5.127$ ($n = 718$). All three values are negative, supporting the above conclusion that the presence of APOE-ε4 alleles can cause the onset of AD not only directly but also indirectly by expanding the ventricle. Moreover, the average NDE is less than the average NIE, confirming the above conclusion that the existence of APOE-ε4 alleles contributes to the onset of

AD mainly through the mediated mechanism.

Table S7: Coefficients and heteroscedasticity robust standard errors (in parentheses) of best linear prediction of GACEs. * $p < 0.05$; ** $p < 0.01$

	$\tau_{c,g}(\mathbf{x})$	$\zeta_{c,g}(0;\mathbf{x})$	$\xi_{c,g}(1;\mathbf{x})$
Constant	−15.336** (0.506)	−7.233** (0.349)	−8.105** (0.311)
Age	0.146** (0.006)	0.079** (0.004)	0.067** (0.004)
Male	−1.152** (0.087)	−0.934** (0.056)	−0.218** (0.052)
Education level	−0.129** (0.018)	−0.041** (0.011)	−0.088** (0.010)
Hispanic or Latino	0.003 (0.414)	−1.767** (0.316)	1.766**(0.228)
White	−0.547** (0.171)	−0.057 (0.130)	−0.490** (0.085)
Married	−0.514** (0.094)	−0.554** (0.065)	0.040 (0.057)

S11 Derivation of Equation (2.4)

Under Assumptions (III) in Section 2.3, identification of the relevant potential outcome in Equation (2.4) is derived as follows:

$$\begin{aligned}
&\mathbb{E}[Y(\mathbf{x}, t', M_t(\mathbf{x}))] = \mathbb{E}[Y(t', M(t))|\mathbf{X} = \mathbf{x}] \\
&= \int \mathbb{E}[Y(t', m)|M(t) = m, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m) \\
&= \int \mathbb{E}[Y(t', m)|T = t, M(t) = m, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m) \\
&= \int \mathbb{E}[Y(t', m)|T = t, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m) \qquad \text{(S11.60)} \\
&= \int \mathbb{E}[Y(t', m)|T = t', \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m) \\
&= \int \mathbb{E}[Y(t', m)|T = t', M(t') = m, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m) \\
&= \int \mathbb{E}[Y|T = t', M = m, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m).
\end{aligned}$$

Bibliography

Arjovsky, M., S. Chintala, and L. Bottou (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019). Nearly-tight

vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research 20*(1), 2285–2301.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*(Nov), 463–482.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT press.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems 27*.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Levin, D. A. and Y. Peres (2017). *Markov Chains and Mixing Times*, Volume 107. American Mathematical Soc.

Lo, V. S. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter 4*(2), 78–86.

Mirza, M. and S. Osindero (2014). Conditional generative adversarial nets. *ArXiv Preprint ArXiv:1411.1784*.

Shen, Z., H. Yang, and S. Zhang (2019). Deep network approximation characterized by number of neurons. *ArXiv Preprint ArXiv:1906.05497*.

Sparapani, R., C. Spanbauer, and R. McCulloch (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: the bart r package. *Journal of Statistical Software 97*, 1–66.

Tabib, S. and D. Larocque (2020). Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics 36*(2), 629–636.

Therneau, T. et al. (2015). A package for survival analysis in s. *R Package Version 2*(7).

Tsybakov, A. (2008). Introduction to Nonparametric Estimation. *Springer Science & Business Media, New York: Springer*.

Yoon, J., J. Jordon, and M. Van Der Schaar (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Zhou, X., Y. Jiao, J. Liu, and J. Huang (2022). A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 1–12.