Statistica Sinica Preprint No: SS-2024-0188	
Title	Conditional Generative Adversarial Network for Individualized Causal Mediation Analysis with Survival Outcome
Manuscript ID	SS-2024-0188
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0188
Complete List of Authors	Cheng Huan, Xinyuan Song and Hongwei Yuan
Corresponding Authors	Xinyuan Song
E-mails	xysong@sta.cuhk.edu.hk
Notice: Accepted author version.	

Statistica Sinica

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

Abstract: Causal mediation analysis aims to investigate the underlying mechanism of how an exposure exerts its effects on the outcome mediated by intermediate variables. However, existing methods for causal mediation analysis in the context of survival models are primarily focused on estimating average causal effects and are difficult to apply to precision medicine. Recently, machine learning has emerged as a promising tool for precisely estimating individualized causal effects without assuming specific model forms. This study proposes a novel method, conditional generative adversarial network (CGAN)-based individualized causal mediation analysis with survival outcomes (CGAN-ICMA-SO), to infer individualized causal effects with survival outcomes based on the CGAN framework. We show that the estimated distribution of the proposed inferential conditional generator converges to the true conditional distribution under mild conditions. Our numerical experiments indicate that CGAN-ICMA-SO surpasses five other state-of-the-art methods. Applying the proposed method to an Alzheimer's disease (AD) Neuroimaging Initiative dataset reveals the individualized direct and indirect effects of the APOE-$\varepsilon 4$ allele on time to AD onset.

Key words and phrases: Causal mediation analysis; CGAN; Individualized causal

effects; Loss function; Time-to-event outcome.

1. Introduction

Mediation analysis is widely used in biomedical research (Sun et al., 2021), epidemiology (VanderWeele and Vansteelandt, 2014), and social-psychological studies to examine how exposure influences outcomes. Exposure affects outcomes both directly and indirectly through mediators, allowing the total effect to be decomposed into direct and indirect effects. Using the counterfactual framework (Imai et al., 2010; Rubin, 2005), this approach, known as causal mediation analysis, is the focus of this paper.

The literature has devoted significant attention to identifying causal effects on survival outcomes. In the context of survival analysis, many existing mediation analyses often rely on linear parametric models, such as linear regression models or linear structural equation models (LSEM) with additive hazards (AH), proportional hazards (PH), or accelerated failure time (AFT) models (VanderWeele, 2011). These methods decompose the average treatment effect into natural direct and indirect effects at the population level. However, these approaches often assume homogeneity across individuals and focus only on average causal mediation effects. In biomedical studies, it is often observed that patients with different characteristics can have varying

disease progressions, even when subjected to the same treatment. The lack of homogeneity in treatment effects has garnered significant attention, mainly when dealing with survival outcomes. Consequently, there has been a significant increase in methods developed to estimate individualized treatment effects (ITE) or conditional average treatment effects (CATE) for survival outcomes. Parametric models like Cox proportional hazards (CoxPH) (Cox, 1972) and the AFT model (Wei, 1992) are commonly used but rely on the unrealistic assumption of correct model specification, limiting their flexibility with high-dimensional data or complex interactions. To address these limitations, machine learning methods, including tree-based and neural network approaches, have emerged. Nonparametric methods like Random Survival Forest (RSF) (Ishwaran et al., 2008) and Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) have been extended for causal survival analysis. RSF has been applied in causal survival forests with weighted bootstrap inference (Cui et al., 2023; Shen et al., 2018), while BART has been adapted to consider survival outcomes in models like Surv-BART (Sparapani et al., 2016) and AFT-BART (Henderson et al., 2020). Additionally, SurvITE (Curth et al., 2021), a deep learning approach, was introduced to characterize diverse treatment effects on survival probabilities. When the total effect of exposure on a survival outcome is heterogeneous, understanding how this heterogeneity arises across causal pathways is crucial. This helps reveal the underlying mechanisms through which heterogeneity in treatment effects

emerges, thereby offering deeper insights into complex biomedical dynamics.

Detecting potential heterogeneity in mediation effects has long been a focal point in psychology, known as moderated mediation analysis (Hayes, 2015; Preacher et al., 2007). Existing methods for estimating the conditional average causal effects or individualized causal effects (ICEs) are primarily tailored for continuous outcomes. For instance, Park and Kaplan (2015) integrated Bayesian inferential approaches with G-computation for mediation analysis in group randomized designs. Qin and Hong (2017) introduced a weighting technique to identify and estimate site-specific mediation effects. Later on, Dyachenko and Allenby (2018) proposed a Bayesian mixture model that combines likelihood functions based on two distinct outcome models. Xue et al. (2022) introduced a novel mediation penalty for high-dimensional data. Qin and Wang (2023) further defined causal conditional effects with moderated mediation effects across subgroups. While these methods have proven beneficial in various scenarios, they exhibit limitations. Approaches like Park and Kaplan (2015) and Xue et al. (2022), heavily lean on the LSEM framework, which may not capture complex real-world systems. Qin and Hong (2017) primarily focused on estimating the population average and between-site variance of causal effects, neglecting subpopulation-specific effects. Qin and Wang (2023) emphasized the importance of accurately specifying parametric mediator and outcome models. Dyachenko and Allenby (2018) introduced a method

that necessitates a predetermined number of subgroups. It is also crucial to note that the above methods have primarily focused on continuous outcomes, highlighting the urgent need for tailored approaches to heterogeneous mediation analysis with time-to-event outcomes. Developing advanced methods to address this gap is essential.

Machine learning has recently garnered significant interest for capturing complex nonlinear relationships without relying on predefined model forms. Some methods (Bica et al., 2020; Chen et al., 2019; Chu et al., 2020; Ge et al., 2020; Yoon et al., 2018) have focused on estimating ITE with continuous outcomes, while others (Chapfuwa et al., 2021; Curth et al., 2021) have addressed ITE with survival outcomes. Additionally, certain methodologies (Huan et al., 2024; Xu et al., 2022) have targeted the estimation of ICEs with continuous outcomes. While these methods effectively estimate ITE and ICEs, to our knowledge, there is a gap in addressing ICEs with survival outcomes. Notably, deep learning has shown promise in survival analysis, such as Chapfuwa et al. (2018)'s adversarial learning for nonparametric time-to-event analysis, though it does not focus on causal inference. Among these methods, Yoon et al. (2018) introduced a Conditional Generative Adversarial Nets (CGAN)-based method called GANITE for inferring ITE. This approach comprises two blocks: a counterfactual imputation block and an ITE block, each containing a generator and a discriminator. Furthermore, there are extensions of GAN-

ITE (Bica et al., 2020; Ge et al., 2020) that introduce techniques to evaluate ITE of different treatment types using GANs and Huan et al. (2024) that extend GANITE to assess ICEs with continuous outcomes. However, none simultaneously address ICEs and survival outcomes. Estimating ICEs, as in Equation (2.4), requires a sampling-based technique to draw samples from the estimated probability distribution of the potential mediator and assess potential outcomes. Leveraging GANs' ability to capture complex nonlinear relationships and sample from probability distributions, we aim to propose a novel method, CGAN-based individualized causal mediation analysis with survival outcome (CGAN-ICMA-SO), to bridge this gap.

The CGAN-ICMA-SO model estimates ICE and explores individualized causal mechanisms for survival outcomes. This method is structured with two key layers: the mediator layer and the outcome layer. Each layer is further divided into two subblocks: the counterfactual block and the inferential block. These subblocks incorporate a generator and a discriminator, enabling precise estimation of causal effects at the individual level. Furthermore, we establish distribution matching estimation and prove the convergence of CGAN-ICMA-SO, ensuring that the estimated distribution from our inferential generator converges to the true conditional distribution under mild conditions. Our proposed method addresses the limitations by capturing complex nonlinear relationships without relying on parametric structures like LSEM, ensuring

accurate modeling of real-world systems. It is versatile, requiring no assumptions about heterogeneity sources or predefined subgroups, making it widely applicable. Additionally, the distribution matching estimation and convergence theory of CGAN-ICMA-SO provide a strong theoretical foundation, ensuring accurate conditional distribution modeling and reliable sampling. These guarantees enhance the method's reliability and validity, supporting its application and further development.

This research was motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The apolipoprotein E-ε4 (APOE-ε4) allele is strongly associated with ventricle expansion, a known contributor to AD development. Our research interest is investigating the causal mechanism linking APOE-ε4 to time-to-AD onset and how this mechanism varies across observable characteristics like age and gender. Traditional methods estimate average causal effects and fail to handle situations where causal effects differ across subgroups defined by observable characteristics, as noted in previous mediation studies (Sun et al., 2021; Zhou and Song, 2021). To address this, we propose CGAN-ICMA-SO, which allows for personalized estimates of ICEs. Our findings reveal that APOE-ε4 accelerates AD onset both directly and indirectly by enlarging the ventricle, with the mediated pathway having a greater impact. Additionally, our method identifies how causal effects vary across characteristics such as age, gender, education, and ethnicity. For instance, we observe

that APOE-ε4's impact on AD onset increases initially and then stabilizes as education level rises, as discussed further in Section 7.

The organization of this paper is as follows. Section 2 presents problem formulation and briefly reviews ICEs with survival outcomes. Section 3 presents the proposed CGAN-ICMA-SO. Section 4 establishes the convergence of CGAN-ICMA-SO, and its implementation is provided in Section 5. Section 6 compares the proposed method with several other approaches by simulation studies. Section 7 applies our method to the ADNI dataset, and Section 8 concludes. The theoretical proofs, other technical details, and parts of numerical results are relegated to the Supplementary Material.

2. Problem Formulation

2.1 Preliminary

Suppose we have a random vector $\mathbf{X} = (X_1, ..., X_{d_x}) \in \mathcal{X} \subset \mathbb{R}^{d_x}$ representing pre-treatment covariates, a binary exposure or treatment indicator $T \in \mathcal{T} := \{0, 1\}$, and another random variable, namely, the mediator $M \in \mathcal{M} \subset \mathbb{R}$. Additionally, we have two random variables: $Y \in \mathcal{Y} \subset \mathbb{R}$ representing the event time and $C \in \mathcal{C} \subset \mathbb{R}$ denoting the noninformative censoring time.

Assume that $(\mathbf{X}, T, M, Y, C) \sim P_{\mathbf{X},T,M,Y,C}$ with marginal distributions such as $(\mathbf{X}, T, M, Y) \sim P_{\mathbf{X},T,M,Y}$, $(\mathbf{X}, T, M) \sim P_{\mathbf{X},T,M}$, $(\mathbf{X}, T) \sim P_{\mathbf{X},T}$ and so forth. Furthermore, we have individual distribution functions for each variable, such as $\mathbf{X} \sim P_{\mathbf{X}}$, $T \sim P_T$, $M \sim P_M$, $Y \sim P_Y$, $C \sim P_C$, etc. We denote the conditional distribution of Y given $(\mathbf{X}, T, M)$ as $P_{Y|\mathbf{X},T,M}$, and

similar notations are used for other conditional distributions. Let $p_{\mathbf{X},T,M,Y}$ be the density function of the distribution $P_{\mathbf{X},T,M,Y}$, and similar notations are used for other distributions.

Let $M(t)$ be a potential mediating variable that represents the value of the mediator if the treatment variable is equals to $t \in \mathcal{T}$ and let $Y(t,m)$ be the potential event time if one receives treatment $t \in \mathcal{T}$ and mediator $m \in \mathcal{M}$. The factual mediator, the factual event time, and the observed time are denoted by $M = M(T)$, $Y = Y(T, M(T))$, and $\widetilde{Y} = \min\{Y(T, M(T)), C\}$, respectively, where T is the factual treatment. In addition, we make the consistency assumption throughout: for any individual $\mathbf{X} = \mathbf{x}$, the potential mediator $M(t)$ is equal to the observed mediator $M = M(T)$ if the individual $\mathbf{X} = \mathbf{x}$ happened to receive treatment level $T = t$; and so do the potential event time. Denote $M(\mathbf{x}, t) := M(t)|\mathbf{X} = \mathbf{x}$ and $Y(\mathbf{x}, t, m) := Y(t, m)|\mathbf{X} = \mathbf{x}$. To emphasize the individualized effects, denote $M_t(\mathbf{x}) := M(\mathbf{x}, t)$ for $t \in \mathcal{T}$, $Y_t(\mathbf{x}, m) := Y(\mathbf{x}, t, m)$, and $Y_{t't}(\mathbf{x}) := Y(\mathbf{x}, t', M_t(\mathbf{x}))$, for $t, t' \in \mathcal{T}$. For any given $\mathbf{X} = \mathbf{x}$, we aim to obtain the distribution of $M_t(\mathbf{x})$, then the value of $\mathbb{E}[Y_{t't}(\mathbf{x})]$ for $t, t' \in \mathcal{T}$, and finally, the ICEs defined in Section 2.3. Denote the failure indicator by $\delta(\mathbf{x}, t, m) = \mathbb{I}\big(Y(\mathbf{x}, t, m) \leq C\big)$. Let $\mathbf{Z}$, $\widehat{\mathbf{Z}}$, $\widetilde{\mathbf{Z}}$ and $\overline{\mathbf{Z}}$ be random vectors independent of $\mathbf{X}$, T, M, Y and each other, with a known distribution $P_{\mathbf{Z}}$, $P_{\widehat{\mathbf{Z}}}$, $P_{\widetilde{\mathbf{Z}}}$ and $P_{\overline{\mathbf{Z}}}$, respectively. For example, we can take $P_{\mathbf{Z}}$ as the standard multivariate normal distribution $N(0, \mathbb{I}_{d_z})$ for a given $d_z \geq 1$.

2.2 General Assumptions

We now introduce several standard assumptions (Imai et al., 2010; Martinussen et al., 2011; Imbens and Rubin, 2015) to identify causal effects.

Assumptions:

(I) [The stable unit treatment value assumption (SUTVA)] a. The units in the study are stable and do not change their behavior based on the treatment assignment of other units. b. There is a well-defined and consistent treatment value for each unit.

(II) [Overlap] $P(T = t|\mathbf{X} = \mathbf{x}) > 0$ and $P(M(t) = m|T = t, \mathbf{X} = \mathbf{x}) > 0$ for all $(t, m, \mathbf{x}) \in \mathcal{T} \times \mathcal{M} \times \mathcal{X}$.

(III) [Unconfoundedness] $Y(t', m) \perp\!\!\!\perp M(t)|T = t, \mathbf{X} = \mathbf{x}$, $\{Y(t', m), M(t)\} \perp\!\!\!\perp T|\mathbf{X} = \mathbf{x}$, for all $(t, t', m, \mathbf{x}) \in \mathcal{T} \times \mathcal{T} \times \mathcal{M} \times \mathcal{X}$.

(IV) [Noninformative Censoring] $C \perp\!\!\!\perp Y|\mathbf{X} = \mathbf{x}, T = t, M = m$, for all $\mathbf{x} \in \mathcal{X}$, $t, \in \{0, 1\}$ and $m \in \mathcal{M}$.

These assumptions are explained in Supplementary Material S1.

2.3 Problems

For dataset $S_n := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = \widetilde{y}_i, \delta = \delta_i\}_{i=1}^n$, we assume that the censoring rate is α_r, where $0 \leq \alpha_r < 1$, and split S_n into two sub-datasets based on the censoring indicator. Let $S_{n_1}^{(1)} := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, Y = y_i, \delta = \delta_i = 1\}_{i=1}^{n_1} \subset \mathcal{X} \times \{0, 1\} \times \mathcal{M} \times \mathcal{Y} \times \{1\}$ be the sub-dataset containing the observed event times ($\delta = 1$) and $S_{n_2}^{(2)} := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M =$

$m_i, C = c_i, \delta = \delta_i = 0\}_{i=n_1+1}^{n} \subset \mathcal{X} \times \{0,1\} \times \mathcal{M} \times \mathcal{C} \times \{0\}$ be the sub-dataset containing the censored observations ($\delta = 0$), where $n = n_1 + n_2$ and $n_2 = \alpha_r n$ with $\alpha_r \in [0, 1)$. We aim to approximate the expectation of event times for a given covariate $\mathbf{X} = \mathbf{x}$ with different treatments, that is, $\mathbf{x} \mapsto \mathbb{E}[Y_{t't}(\mathbf{x})]$, for $t', t \in \{0, 1\}$, i.e., $\mathbb{E}[Y(\mathbf{x}, 0, M_0(\mathbf{x}))]$, $\mathbb{E}[Y(\mathbf{x}, 0, M_1(\mathbf{x}))]$, $\mathbb{E}[Y(\mathbf{x}, 1, M_0(\mathbf{x}))]$, and $\mathbb{E}[Y(\mathbf{x}, 1, M_1(\mathbf{x}))]$. Then, ICEs can be approximated by comparing the expected event times under different treatment and mediator combinations. Now, we introduce ICEs (Huang and Yang, 2017; Imai et al., 2010; Royston and Parmar, 2011). First, we define the individualized natural indirect effect (NIE) and natural direct effect (NDE) of treatment as

$$\text{NIE} = \xi(t; \mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, t, M_1(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, t, M_0(\mathbf{x}))], \tag{2.1}$$

$$\text{NDE} = \zeta(t; \mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, 1, M_t(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, 0, M_t(\mathbf{x}))], \tag{2.2}$$

for $t \in \{0, 1\}$ and $\mathbf{x} \in \mathcal{X}$. Given the notations introduced in Section 2.1, the expectations in (2.1) and (2.2) are conditional expectations given $\mathbf{x}$, e.g., $\mathbb{E}[Y(\mathbf{x}, t, M_1(\mathbf{x}))] = \mathbb{E}[Y(t, M(1))|\mathbf{X} = \mathbf{x}]$. Then, the individualized total effect (TE) of treatment can be decomposed as

$$\tau(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, 1, M_1(\mathbf{x}))] - \mathbb{E}[Y(\mathbf{x}, 0, M_0(\mathbf{x}))] = \frac{1}{2}\sum_{t=0}^{1}\{\xi(t; \mathbf{x}) + \zeta(t; \mathbf{x})\}, \tag{2.3}$$

for $\mathbf{x} \in \mathcal{X}$. In ADNI data analysis in Section 7, the individualized NDE or NIE denotes the effect of carrying APOE-ε4 alleles on the time to AD onset without through or through ventricle expansion, respectively. The individualized TE represents the total effect of carrying APOE-ε4 alleles on AD onset.

To demonstrate the identifiability of ICEs, we examine the identification of the relevant potential outcome as follows. By Assumption (III), we have

$$\mathbb{E}[Y(\mathbf{x}, t', M_t(\mathbf{x}))] = \mathbb{E}[Y(t', M(t))|\mathbf{X} = \mathbf{x}] = \int \mathbb{E}[Y|T = t', M = m, \mathbf{X} = \mathbf{x}] dP_{M(t)|\mathbf{X}=\mathbf{x}}(m). \tag{2.4}$$

Thus, ICEs can be identified through the expected potential outcomes as long as the distribution of the potential mediator, $P_{M(\mathbf{x},t)}$, can be estimated from the observed data. Detailed derivation of Equation (2.4) is provided in Supplementary Material S11. Next, we present the problem formulation.

Problem in M Layer: Find a deterministic function, called inference function for mediator M, $\mathbf{I_M} : (\widehat{\mathbf{z}}, \mathbf{x}) \in \mathbb{R}^{d_z} \times \mathcal{X} \mapsto \mathbf{I_M}(\widehat{\mathbf{z}}, \mathbf{x}) = (I^{(0)}_{\mathbf{M}}(\widehat{\mathbf{z}}, \mathbf{x}), I^{(1)}_{\mathbf{M}}(\widehat{\mathbf{z}}, \mathbf{x})) \in \mathcal{M} \times \mathcal{M}$, such that

$$\mathbf{I_M}(\widehat{\mathbf{Z}}, \mathbf{x}) \sim P_{M|\mathbf{X}=\mathbf{x},T=0} \otimes P_{M|\mathbf{X}=\mathbf{x},T=1}, \ \mathbf{x} \in \mathcal{X}, \tag{2.5}$$

which means that $I^{(0)}_{\mathbf{M}}(\widehat{\mathbf{Z}}, \mathbf{x}) \sim P_{M|\mathbf{X}=\mathbf{x},T=0}$ and $I^{(1)}_{\mathbf{M}}(\widehat{\mathbf{Z}}, \mathbf{x}) \sim P_{M|\mathbf{X}=\mathbf{x},T=1}$. Thus, for $\mathbf{x} \in \mathcal{X}$, to sample from $P_{M|\mathbf{X}=\mathbf{x},T=0}$ and $P_{M|\mathbf{X}=\mathbf{x},T=1}$, we can first sample a $\widehat{\mathbf{z}} \sim P_{\widehat{\mathbf{Z}}}$ and then calculate $\mathbf{I_M}(\widehat{\mathbf{z}}, \mathbf{x})$. The resulting value $I^{(0)}_{\mathbf{M}}(\widehat{\mathbf{z}}, \mathbf{x})$ is a sample from $P_{M|\mathbf{X}=\mathbf{x},T=0}$, and $I^{(1)}_{\mathbf{M}}(\widehat{\mathbf{z}}, \mathbf{x})$ is a sample from $P_{M|\mathbf{X}=\mathbf{x},T=1}$.

Problem in Y Layer: Find a deterministic function, called inference function for the event time Y, $\mathbf{I_Y} : (\overline{\mathbf{z}}, \mathbf{x}, m) \in \mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M} \mapsto \mathbf{I_Y}(\overline{\mathbf{z}}, \mathbf{x}, m) = (I^{(0)}_{\mathbf{Y}}(\overline{\mathbf{z}}, \mathbf{x}, m), I^{(1)}_{\mathbf{Y}}(\overline{\mathbf{z}}, \mathbf{x}, m)) \in \mathcal{Y} \times \mathcal{Y}$, such that

$$\mathbf{I_Y}(\overline{\mathbf{Z}}, \mathbf{x}, m) \sim P_{Y|\mathbf{X}=\mathbf{x},T=0,M=m} \otimes P_{Y|\mathbf{X}=\mathbf{x},T=1,M=m}, \ \mathbf{x} \in \mathcal{X}, \ m \in \mathcal{M}. \tag{2.6}$$

Once the above problems are resolved, we calculate $\mathbb{E}[Y_{t't}(\mathbf{x})]$ as follows:

$$\mathbb{E}[Y_{t't}(\mathbf{x})] = \mathbb{E}_{\widehat{\mathbf{Z}}\sim P_{\widehat{\mathbf{Z}}},\overline{\mathbf{Z}}\sim P_{\overline{\mathbf{Z}}}}[I_{\mathbf{Y}}^{(t')}(\overline{\mathbf{Z}},\mathbf{x},I_{\mathbf{M}}^{(t)}(\widehat{\mathbf{Z}},\mathbf{x}))]$$

for all $\mathbf{x}\in\mathcal{X}$. Subsequently, we can utilize these estimates to obtain ICEs with survival time. We can use Monte Carlo approximations to empirically evaluate the integral (2.4). Specifically, we sample $\widehat{n}$ samples from $\widehat{\mathbf{Z}}\sim P_{\widehat{\mathbf{Z}}}$ and $\overline{n}$ samples from $\overline{\mathbf{Z}}\sim P_{\overline{\mathbf{Z}}}$, denoted as $\widehat{\mathbf{z}}_1,\widehat{\mathbf{z}}_2,\ldots,\widehat{\mathbf{z}}_{\widehat{n}}$ and $\overline{\mathbf{z}}_1,\overline{\mathbf{z}}_2,\ldots,\overline{\mathbf{z}}_{\overline{n}}$. Then, NIE and NDE in (2.1) and (2.2) can be estimated by

$$\xi(t;\mathbf{x})\approx\frac{1}{\overline{n}\times\widehat{n}}\left(\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(t)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_i,\mathbf{x}))-\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(t)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_i,\mathbf{x}))\right), \quad (2.7)$$

$$\zeta(t;\mathbf{x})\approx\frac{1}{\overline{n}\times\widehat{n}}\left(\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_i,\mathbf{x}))-\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_i,\mathbf{x}))\right), \quad (2.8)$$

and TE in (2.3) can be estimated by

$$\tau(\mathbf{x})\approx\frac{1}{\overline{n}\times\widehat{n}}\left(\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_i,\mathbf{x}))-\sum_{j=1}^{\overline{n}}\sum_{i=1}^{\widehat{n}}I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_j,\mathbf{x},I_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_i,\mathbf{x}))\right). \quad (2.9)$$

3. Method

We propose a new architecture of CGAN-ICMA-SO, depicted in Figure 1, to address the mediation process, censoring time, and identification of the relevant potential outcome. This architecture consists of a mediator layer and an outcome layer. Each layer consists of two subblocks: a counterfactual block and an inferential block. The inferential generators in the mediator and outcome layers are designed to generate samples that match the target distributions specified in Equations (2.5) and (2.6), respectively.

To streamline the presentation, we present the detailed technical description of the mediator layer in the Supplementary Material because it shares a

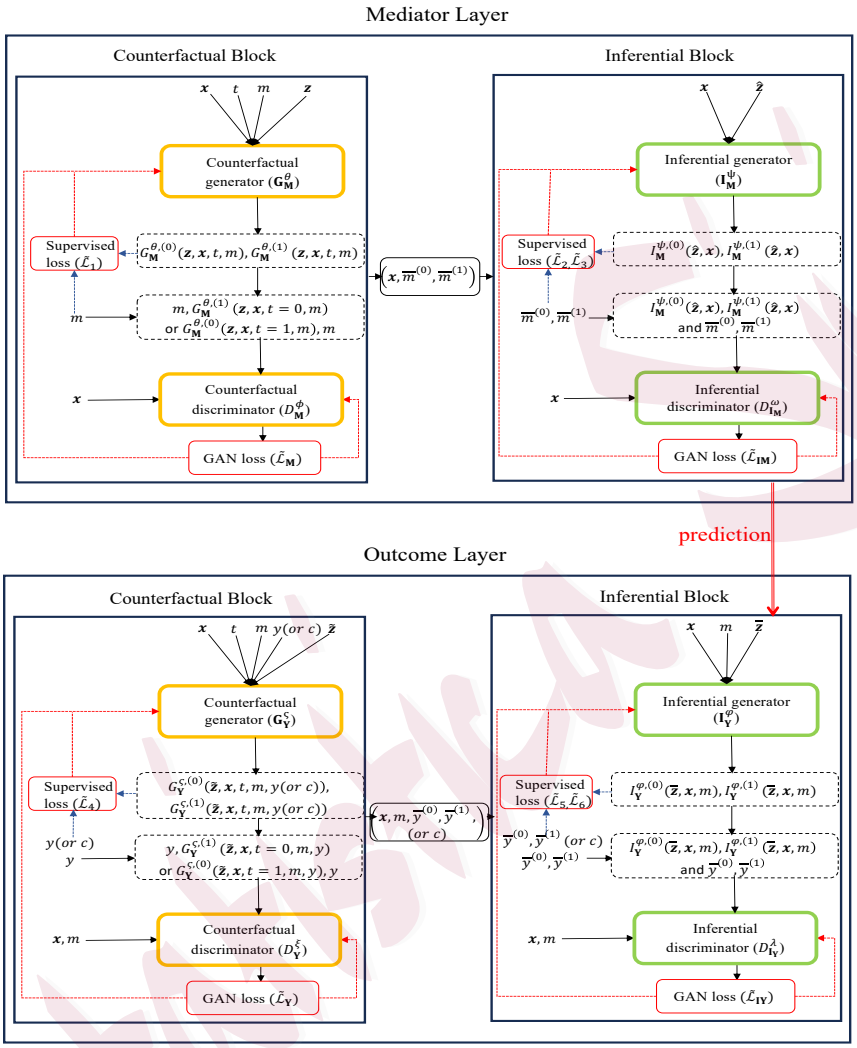

Figure 1: Architecture of CGAN-ICMA-SO: $(\overline{m}^{(0)}, \overline{m}^{(1)})$ is sampled from $\widehat{\mathbf{G}}_{\mathbf{M}}$ after $\widehat{\mathbf{G}}_{\mathbf{M}}$ has been fully trained and $(\overline{y}^{(0)}, \overline{y}^{(1)})$ is sampled from $\widehat{\mathbf{G}}_{\mathbf{Y}}$ after $\widehat{\mathbf{G}}_{\mathbf{Y}}$ has been fully trained. $\mathbf{G}_{\mathbf{M}}^{\boldsymbol{\theta}}$, $D_{\mathbf{M}}^{\boldsymbol{\phi}}$, $\mathbf{G}_{\mathbf{Y}}^{\boldsymbol{\zeta}}$, $D_{\mathbf{Y}}^{\boldsymbol{\xi}}$, $D_{\mathbf{I}_{\mathbf{M}}}^{\boldsymbol{\omega}}$, and $D_{\mathbf{I}_{\mathbf{Y}}}^{\boldsymbol{\lambda}}$ are only operating during training, whereas $\mathbf{I}_{\mathbf{M}}^{\boldsymbol{\psi}}$ and $\mathbf{I}_{\mathbf{Y}}^{\boldsymbol{\varphi}}$ operate both during training and at run-time. $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\psi}$, and $\boldsymbol{\varphi}$ represent the trainable parameters in the FNN network. The generic notations $\mathbf{G}$ and D represent the generator and discriminator in CGAN, respectively.

structural analogy with the outcome layer but operates under a simpler setting. Specifically, the mediator layer adapts the outcome layer's framework through corresponding notational adjustments: replacing the pair (X, M) with only X as input and substituting the survival outcome Y with the fully observed mediator M. This eliminates the need to handle censoring mechanisms within the mediator layer, thereby simplifying both the design of loss function and theoretical analysis. Moreover, this prioritization allows us to focus on the methodological novelty of handling survival outcomes in the outcome layer, which introduces additional complexities such as censoring adjustments. Interested readers may refer to Supplementary Material S2.1 for the details of the mediator layer. In the mediator layer, we utilize the feedforward neural network (FNN) for estimation, deriving the estimated conditional generators and discriminators in the counterfactual and inferential blocks: $\widehat{\mathbf{G}}_{\mathbf{M}}, \widehat{D}_{\mathbf{M}}, \widehat{\mathbf{I}}_{\mathbf{M}}$, and $\widehat{D}_{\mathbf{I}_{\mathbf{M}}}$.

3.1 Outcome Layer

Counterfactual Block: it consists of a generator and a discriminator. The generator in the counterfactual outcome block, denoted as $\mathbf{G}_{\mathbf{Y}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \{0,1\} \times \mathcal{M} \times \mathcal{Y} \mapsto \mathcal{Y} \times \mathcal{Y}$, takes the covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$), binary treatment variable t (where $T = t$), factual mediator m (where $M = m$), factual outcome y (where $Y = y$), and some noise $\widetilde{\mathbf{Z}}$ as inputs. It generates the com-

plete outcome vector $\mathbf{G_Y}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y) = \big(G_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y), G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y)\big)$. We can represent the random variable induced by $\mathbf{G_Y}$ as $\mathbf{G_Y}(\widetilde{\mathbf{Z}}, \mathbf{X}, T, M, Y) = \big(G_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T, M, Y), G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T, M, Y)\big)$. The discriminator takes $\mathbf{x}, m, (1-t)y + tG_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y)$, and $ty + (1-t)G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y)$ as inputs. It outputs a scalar representing the probability that the last input $ty + (1-t)G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y)$ corresponds to the factual outcome rather than the counterfactual outcome. This setup allows us to generate the complete outcome vector by incorporating the covariates, treatment, factual mediator, factual outcome and noise into the generator and then leveraging the discriminator to distinguish between the factual and counterfactual outcomes. The loss function associated with this setup is:

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

At the population level, the target conditional generator $\mathbf{G}_{\mathbf{Y}}^*$ and discriminator $D_{\mathbf{Y}}^*$ are defined as $(\mathbf{G}_{\mathbf{Y}}^*, D_{\mathbf{Y}}^*) := \operatorname{argmin}_{\mathbf{G_Y}} \operatorname{argmax}_{D_{\mathbf{Y}}} \mathcal{L}_{\mathbf{Y}}(\mathbf{G_Y}, D_{\mathbf{Y}})$, and denote $\mathbf{G}_{\mathbf{Y}}^* = (G_{\mathbf{Y}}^{*,(0)}, G_{\mathbf{Y}}^{*,(1)})$.

Empirical Loss Function of Counterfactual Block: for the dataset $S_n = \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = \widetilde{y}_i, \delta = \delta_i\}_{i=1}^n = S_{n_1}^{(1)} \cup S_{n_2}^{(2)}$ and $\{\widetilde{\mathbf{Z}} = \widetilde{\mathbf{z}}_i\}_{i=1}^n$ independently generated from $P_{\widetilde{\mathbf{Z}}}$, where $S_{n_1}^{(1)} = \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = y_i, \delta = \delta_i = 1\}_{i=1}^{n_1}$ and $S_{n_2}^{(2)} = \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = c_i, \delta =$

$\delta_i = 0\}_{i=n_1+1}^{n}$, we define the sample set $S_n^Y := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \widetilde{Y} = \widetilde{y}_i, \delta = \delta_i, \widetilde{\mathbf{Z}} = \widetilde{\mathbf{z}}_i\}_{i=1}^{n}$ and use it to train $\widehat{\mathbf{G}}_{\mathbf{Y}}$ in the counterfactual block. We consider the following empirical version of $\mathcal{L}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}, D_{\mathbf{Y}})$:

$$\widetilde{\mathcal{L}}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}, D_{\mathbf{Y}}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\Big\{ t_i \log D_{\mathbf{Y}}\Big(\mathbf{x}_i, m_i, (1-t_i)y_i + t_i G_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i), t_i y_i + (1-t_i) G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i)\Big) + (1-t_i)\log\Big[1 - D_{\mathbf{Y}}\Big(\mathbf{x}_i, m_i, (1-t_i)y_i + t_i G_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i), t_i y_i + (1-t_i) G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i)\Big)\Big]\Big\}.$$

We introduce the supervised loss (Chapfuwa et al., 2018) to ensure $G_{\mathbf{Y}}^{(t)}(\widetilde{\mathbf{z}}, \mathbf{x}, t, m, y) = y$ when $\delta = 1$ and $G_{\mathbf{Y}}^{(t)}(\widetilde{\mathbf{z}}, \mathbf{x}, t, m, c) \geq c$ when $\delta = 0$: $\widetilde{\mathcal{L}}_4(\mathbf{G}_{\mathbf{Y}}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\Big| G_{\mathbf{Y}}^{(t_i)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, y_i) - y_i\Big|^2 + \frac{1}{n_2}\sum_{i=n_1+1}^{n}\Big\{\max\{0, c_i - G_{\mathbf{Y}}^{(t_i)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, t_i, m_i, c_i)\}\Big\}^2$. Define an empirical objective function for a supervised parameter $\alpha_4 \geq 0$:

$$\widehat{\mathcal{L}}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}, D_{\mathbf{Y}}) := \widetilde{\mathcal{L}}_{\mathbf{Y}}(\mathbf{G}_{\mathbf{Y}}, D_{\mathbf{Y}}) + \alpha_4 \widetilde{\mathcal{L}}_4(\mathbf{G}_{\mathbf{Y}}). \tag{3.10}$$

We use FNN to estimate $\mathbf{G}_{\mathbf{Y}}$, denoted as $\widehat{\mathbf{G}}_{\mathbf{Y}}$, based on the empirical objective function (3.10). See details in Supplementary Material S2.5.

Inferential Block: After training the counterfactual outcome block, we obtain the vector $\Big((1-t)y + tG_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y), ty + (1-t)G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y)\Big)$. Next, we pass this vector, along with the given covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$) and factual mediator m (where $M = m$), into the inferential outcome block to obtain the outcome vector, denoted as $\mathbf{I}_{\mathbf{Y}}(\overline{\mathbf{Z}}, \mathbf{x}, m) = (I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}}, \mathbf{x}, m), I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}}, \mathbf{x}, m))$, the generator $\mathbf{I}_{\mathbf{Y}} : \mathbb{R}^{d_z} \times \mathcal{X} \times \mathcal{M} \mapsto \mathcal{Y} \times \mathcal{Y}$ takes the covariates $\mathbf{x}$ (where $\mathbf{X} = \mathbf{x}$), factual mediators m (where $M = m$), and some noise $\overline{\mathbf{Z}}$ as inputs, without relying on factual outcome and treatment. The output is $\mathbf{I}_{\mathbf{Y}}(\overline{\mathbf{Z}}, \mathbf{x}, m)$, and we can represent the random variable induced by $\mathbf{I}_{\mathbf{Y}}$ as $\mathbf{I}_{\mathbf{Y}}(\overline{\mathbf{Z}}, \mathbf{X}, M) = (I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}}, \mathbf{X}, M), I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}}, \mathbf{X}, M))$. The discriminator takes either $(\mathbf{x}, m, (1-t)y + tG_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{Z}},$

$\mathbf{x}, t, m, y), ty+(1-t)G_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{Z}}, \mathbf{x}, t, m, y))$ or $\big(\mathbf{x}, m, \mathbf{I}_{\mathbf{Y}}(\overline{\mathbf{Z}}, \mathbf{x}, m)\big)$ as inputs. By employing this architecture, we can generate the complete outcome vector by integrating the covariates, factual mediator, and noise into the generator. The discriminator helps distinguish between the counterfactual and inferred complete outcome vectors. We utilize the classical CGAN loss, denoted by

$$\mathcal{L}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}}) := \mathbb{E}_{\widetilde{\mathbf{q}} \sim P_{\widetilde{\mathbf{Q}}}}[\log D_{\mathbf{I}_{\mathbf{Y}}}(\widetilde{\mathbf{q}})] + \mathbb{E}_{(\mathbf{x},m)\sim P_{\mathbf{X},M}} \mathbb{E}_{\overline{\mathbf{z}} \sim P_{\overline{\mathbf{Z}}}} \log \Big[1 - D_{\mathbf{I}_{\mathbf{Y}}}\big(\mathbf{x}, m, I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}, \mathbf{x}, m), I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}, \mathbf{x}, m)\big)\Big],$$

where $P_{\widetilde{\mathbf{Q}}}$ is the joint distribution of $\Big(\mathbf{X}, M, G_{\mathbf{Y}}^{*,(0)}(\widetilde{\mathbf{Z}}, \mathbf{X}, T=1, M, Y_1(\mathbf{X}, M)), Y_1(\mathbf{X}, M)\Big)$. Define $(\mathbf{I}_{\mathbf{Y}}^*, D_{\mathbf{I}_{\mathbf{Y}}}^*) := \text{argmin}_{\mathbf{I}_{\mathbf{Y}}} \text{argmax}_{D_{\mathbf{I}_{\mathbf{Y}}}} \mathcal{L}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}})$, where $\mathbb{L}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}) := \sup_{D_{\mathbf{I}_{\mathbf{Y}}}} \mathcal{L}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}})$, and denote $\mathbf{I}_{\mathbf{Y}}^* = (I_{\mathbf{Y}}^{*,(0)}, I_{\mathbf{Y}}^{*,(1)})$.

By the Lemmas on distribution matching shown in Supplementary Material S2.2, we can conclude 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 $I_{\mathbf{Y}}^{*,(1)}(\overline{\mathbf{Z}}, \mathbf{X}, M) \sim Y_1(\mathbf{X}, M) \sim P_{Y|\mathbf{X},T=1,M}$.

Empirical Loss Function of Inferential Block: given the sample set S_n^Y and $\{\overline{\mathbf{Z}} = \overline{\mathbf{z}}_i\}_{i=1}^n$ independently generated from $P_{\overline{\mathbf{Z}}}$, we use $\widehat{\mathbf{G}}_{\mathbf{Y}}$ obtained above to define another sample set $S_n^{IY} := \{(\mathbf{x}_i, t_i, m_i, \overline{y}_i^{(0)}, \overline{y}_i^{(1)}, \overline{\mathbf{z}}_i)\}_{i=1}^{n_1} \cup \{\mathbf{x}_i, t_i, m_i, c_i, \overline{\mathbf{z}}_i\}_{i=n_1+1}^n$, where $(\overline{y}_i^{(0)}, \overline{y}_i^{(1)}) = t_i(\widehat{G}_{\mathbf{Y}}^{(0)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=1, m_i, y_i), y_i) + (1-t_i)(y_i, \widehat{G}_{\mathbf{Y}}^{(1)}(\widetilde{\mathbf{z}}_i, \mathbf{x}_i, T=0, m_i, y_i))$, to train the estimated conditional generator $\widehat{\mathbf{I}}_{\mathbf{Y}}$. Consider the following empirical version of $\mathcal{L}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}})$:

$$\widetilde{\mathcal{L}}_{\mathbf{IY}}(\mathbf{I}_{\mathbf{Y}}, D_{\mathbf{I}_{\mathbf{Y}}}; \widehat{\mathbf{G}}_{\mathbf{Y}}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \Big\{ \log D_{\mathbf{I}_{\mathbf{Y}}}\Big(\mathbf{x}_i, m_i, \overline{y}_i^{(0)}, \overline{y}_i^{(1)}\Big) + \log \Big[1 - D_{\mathbf{I}_{\mathbf{Y}}}\Big(\mathbf{x}_i, m_i, I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i), I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_i, \mathbf{x}_i, m_i)\Big)\Big] \Big\}.$$

To optimal the performance with respect to $\mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}} \Big\{ \Big| \mathbb{E}\big[Y_1(\mathbf{x}, m) - Y_0(\mathbf{x}, m)\big] - \mathbb{E}\big[I_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}}, \mathbf{x}, m) - I_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}}, \mathbf{x}, m)\big] \Big|^2 \Big\}$, we additionally introduce a supervised loss:

$\widetilde{\mathcal{L}}_5(\mathbf{I_Y};\widehat{\mathbf{G}}_\mathbf{Y}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left|\left(\overline{y}_i^{(0)}-\overline{y}_i^{(1)}\right)-\left(I_\mathbf{Y}^{(0)}(\overline{\mathbf{z}}_i,\mathbf{x}_i,m_i)-I_\mathbf{Y}^{(1)}(\overline{\mathbf{z}}_i,\mathbf{x}_i,m_i)\right)\right|^2$. Furthermore, we introduce another supervised loss (Chapfuwa et al., 2018) to ensure that $I_\mathbf{Y}^{(t)}(\overline{\mathbf{z}},\mathbf{x},m) = y$ when $\delta = 1$ and $I_\mathbf{Y}^{(t)}(\overline{\mathbf{z}},\mathbf{x},m) \geq c$ when $\delta = 0$: $\widetilde{\mathcal{L}}_6(\mathbf{I_Y}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left|I_\mathbf{Y}^{(t_i)}(\overline{\mathbf{z}}_i,\mathbf{x}_i,m_i)-y_i\right|^2+\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left\{\max\{0,c_i-I_\mathbf{Y}^{(t_i)}(\overline{\mathbf{z}}_i,\mathbf{x}_i,m_i)\}\right\}^2$. Define the following empirical objective function, for a supervised parameter α_5, $\alpha_6 \geq 0$,

$$\widehat{\mathcal{L}}_{\mathbf{IY}}(\mathbf{I_Y},D_{\mathbf{I_Y}};\widehat{\mathbf{G}}_\mathbf{Y}) := \widetilde{\mathcal{L}}_{\mathbf{IY}}(\mathbf{I_Y},D_{\mathbf{I_Y}};\widehat{\mathbf{G}}_\mathbf{Y}) + \alpha_5\widetilde{\mathcal{L}}_5(\mathbf{I_Y};\widehat{\mathbf{G}}_\mathbf{Y}) + \alpha_6\widetilde{\mathcal{L}}_6(\mathbf{I_Y}). \tag{3.11}$$

We also use FNN to estimate $\mathbf{I_Y}$, denoted as $\widehat{\mathbf{I}}_\mathbf{Y}$, based on (3.11). See details in Supplementary Material S2.6.

4. Convergence of CGAN-ICMA-SO

This section proves that the distribution of $(\mathbf{X},\widehat{I}_\mathbf{M}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X}))$ (and respectively, $(\mathbf{X},\widehat{I}_\mathbf{M}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X}))$) converges in the total variation norm to the distribution of $(\mathbf{X},M_1(\mathbf{X}))$ (and respectively, $(\mathbf{X},M_0(\mathbf{X}))$) as sample size n goes to infinity within the mediator layer. Furthermore, we show that the distribution of $(\mathbf{X},M,\widehat{I}_\mathbf{Y}^{(1)}(\overline{\mathbf{Z}},\mathbf{X},M))$ (and respectively, $(\mathbf{X},M,\widehat{I}_\mathbf{Y}^{(0)}(\overline{\mathbf{Z}},\mathbf{X},M))$) converges in the same norm as the previous layer to the distribution of $(\mathbf{X},M,Y_1(\mathbf{X},M))$ (and respectively, $(\mathbf{X},M,Y_0(\mathbf{X},M))$) as sample size n goes to infinity within the outcome layer. The regularity conditions and the proofs of theoretical results are given in Supplementary Material S3 and S4, respectively.

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

mediator layer,

$$\mathbb{E}_{S_n^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^n} \| p_{\mathbf{X},M_1(\mathbf{X})} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1} \to 0, \quad as \quad n \to \infty, \tag{4.12}$$

$$\mathbb{E}_{S_n^M \cup \{\widehat{\mathbf{z}}_i\}_{i=1}^n} \| p_{\mathbf{X},M_0(\mathbf{X})} - p_{\mathbf{X},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{Z}},\mathbf{X})} \|_{L^1} \to 0, \quad as \quad n \to \infty, \tag{4.13}$$

where the sample set $S_n^M := \{\mathbf{X} = \mathbf{x}_i, T = t_i, M = m_i, \mathbf{Z} = \mathbf{z}_i\}_{i=1}^n$.

Theorem 2. *Under the assumptions (C.1)–(C.4) and (D.1)–(D.4), then, in outcome layer*

$$\mathbb{E}_{S_n^Y \cup \{\overline{\mathbf{z}}_i\}_{i=1}^n} \| p_{\mathbf{X},M,Y_1(\mathbf{X},M)} - p_{\mathbf{X},M,\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{Z}},\mathbf{X},M)} \|_{L^1} \to 0, \quad as \quad n_1 \to \infty, \tag{4.14}$$

$$\mathbb{E}_{S_n^Y \cup \{\overline{\mathbf{z}}_i\}_{i=1}^n} \| p_{\mathbf{X},M,Y_0(\mathbf{X},M)} - p_{\mathbf{X},M,\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{Z}},\mathbf{X},M)} \|_{L^1} \to 0, \quad as \quad n_1 \to \infty. \tag{4.15}$$

5. Implementation of CGAN-ICMA-SO

We implement CGAN-ICMA-SO using the ReLU as the activation function for training the generators and discriminators. For a detailed description of the training process, please refer to Supplementary Materials S5.

Once the trained model is obtained, we can use the generators $\widehat{\mathbf{I}}_{\mathbf{M}}$ and $\widehat{\mathbf{I}}_{\mathbf{Y}}$ to estimate $\xi(t;\mathbf{x}_e), \zeta(t;\mathbf{x}_e)$, and $\tau(\mathbf{x}_e)$ based on (2.7), (2.8), and (2.9) given the covariates $\mathbf{x}_e$ with $\mathbf{X} = \mathbf{x}_e$. Specifically, we sample $\{\widehat{\mathbf{z}}_1, \widehat{\mathbf{z}}_2, \ldots, \widehat{\mathbf{z}}_{\widehat{n}}\}$ from $\widehat{\mathbf{Z}} \sim P_{\widehat{\mathbf{Z}}}$ and $\{\overline{\mathbf{z}}_1, \overline{\mathbf{z}}_2, \ldots, \overline{\mathbf{z}}_{\overline{n}}\}$ samples from $\overline{\mathbf{Z}} \sim P_{\overline{\mathbf{Z}}}$, we first feed $\mathbf{x}_e$ and $\{\widehat{\mathbf{z}}_i\}_{i=1}^{\widehat{n}}$ into the inferential mediator generator $(\widehat{\mathbf{I}}_{\mathbf{M}})$ to predict $\{\widehat{\mathbf{I}}_{\mathbf{M}}(\widehat{\mathbf{z}}_i, \mathbf{x}_e)\}_{i=1}^{\widehat{n}} = \left\{\left(\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_i, \mathbf{x}_e), \widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_i, \mathbf{x}_e)\right)\right\}_{i=1}^{\widehat{n}}$, Next, we use the different component of $\{\widehat{\mathbf{I}}_{\mathbf{M}}(\widehat{\mathbf{z}}_i, \mathbf{x}_e)\}_{i=1}^{\widehat{n}}$ with $\mathbf{x}_e$ and noise $\{\overline{\mathbf{z}}_i\}_{i=1}^{\overline{n}}$ as the inputs and feed it into the inferential outcome block $(\widehat{\mathbf{I}}_{\mathbf{Y}})$ to generate outcome samples $\widehat{\mathbf{I}}_{\mathbf{Y}}(\overline{\mathbf{z}}_j, \mathbf{x}_e, \widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_i, \mathbf{x}_e)) =$

$\left(\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_j, \mathbf{x}_e, \widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_i, \mathbf{x}_e)), \widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_j, \mathbf{x}_e, \widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_i, \mathbf{x}_e))\right)$ for $t \in \{0,1\}, i \in \{1,\ldots,\widehat{n}\}$, and $j \in \{1,\ldots,\overline{n}\}$. As a result, using these outcome samples, we can obtain the estimation of $\xi(t;\mathbf{x}_e), \zeta(t;\mathbf{x}_e)$, and $\tau(\mathbf{x}_e)$ based on (2.7), (2.8), and (2.9).

6. Simulation study

This section presents simulation studies to evaluate the empirical performance of the proposed method in estimating ICEs with survival time and compare our procedure to five other approaches: linear regression combined with AFT interaction model (LR+AFT), LR combined with two AFT interaction model (LR+2AFT), interaction linear regression combined with another interaction AFT model (ILR+IAFT), random forest combined with random survival forest (RF+RSF), and Bayesian additive regression trees (BART) model. See details in Supplementary Material S6. To evaluate the performance, we utilize three metrics: $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$, $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$, and $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$. See details in Supplementary Material S7. A small value of performance metrics means an accurate estimate. Table S1 in Supplementary Material S8 summarizes the hyperparameters in the network for this simulation. The model setting is as follows:

$$M(\mathbf{x},t) = 0.2 + 2|x_2| + 0.5x_5^2 + 0.1x_6 + t(x_3 + 0.5x_4x_6)^3 + \epsilon_1,$$

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

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

where $M(\mathbf{x},t)$ is a mediator; $Y(\mathbf{x},t,m_t(\mathbf{x}))$ is an event time; C is a censoring

time following $U(0, a)$, $U(0, a)$ is the uniform distribution on $[0, a]$, a is set to yield the censoring rate (CR) of approximately 30% or 50%; $\widetilde{Y}(\mathbf{x}, t, m_t(\mathbf{x}))$ is the observed time, δ is the indicator variable; ϵ_1 and ϵ_2 are independent error terms following $N(0, 0.25)$; $\mathbf{x} = (x_1, \ldots, x_{10})$ is a 10-dimensional covariate vector with $x_3 \sim U(-1, 1)$, $x_4 \sim B(0.4)$, and $x_j \sim N(1.25, 0.4)$ for the rest, $B(0.4)$ is the Bernoulli distribution with a success probability of 0.4; the distribution of treatment t is $P(t = 1) \approx 0.5$.

To save space, we provide the details of the simulation implementation and results in Supplementary Material S9. The results indicate that our method consistently outperforms the other five methods with the smallest values for the averaged $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{TE}}}}$, $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NDE}}}}$, and $\sqrt{\hat{\epsilon}_{\text{PEHE}_{\text{NIE}}}}$.

7. Application: ADNI dataset

This section applies the proposed method to the ADNI dataset to further confirm its utility reflected by the simulation studies in estimating ICEs with survival outcomes. The five other methods are also applied to this dataset for comparison. The ADNI-1 recruited approximately 800 subjects between 55 and 80 and had some extensions afterward. Detailed information about ADNI can be found in the official website: http://adni.loni.usc.edu/. Biomarkers collected in the ADNI study include the number of APOE-ϵ4 alleles, ventricle volume, visit date, and several pre-treatment variables, such as age, gen-

der, education level, ethnicity, race, and marital status. We exclude missing data and focus on 718 patients from the ADNI-1 dataset with mild cognitive impairment (MCI) to explore the underlying causal mechanism of the neurodegenerative progression to dementia and possible heterogeneity.

Previous studies have shown that APOE-ϵ4 alleles are strongly associated with ventricle expansion, further affecting AD development (Roussotte et al., 2014; Sun et al., 2021; Thompson et al., 2004). Thus, the individualized causal mediation analysis for the survival outcome is implemented as follows. The treatment (T) is defined as the presence of APOE-ϵ4 alleles (1 = presence). The difference in the proportion of ventricle volume in the whole brain between the 12th month and the baseline is defined as the observed mediator (M), which is standardized before analysis. The observed time ($\widetilde{Y} = \min(Y, C)$) is the duration from the baseline to the date of the first diagnosis of AD or the date of the last visit, whichever came first, with a CR of 56.7%. Age (x_1), gender (x_2, 1 = male), education level (x_3), ethnicity (x_4, 1 = Hispanic or Latino), race (x_5, 1 = white), and marital status (x_6, 1 = has been married) are included as the baseline covariates.

The original dataset S of 718 samples is randomly split into ten mutually exclusive folds $S_1, \ldots, S_{10}$ of approximately equal size: 72 samples in each of the first nine folds and 70 in the last. At each round $k \in \{1, \ldots, 10\}$, we

train our model on $S\backslash S_k$ with 10,000 iterations and the same set of network hyperparameters as in the simulation. The trained model is then used to predict the remaining samples in the testing set S_k. For robustness, we repeat the analysis 100 times and report the average values of the predicted values. Thus, we can make predictions for the whole dataset after ten rounds.

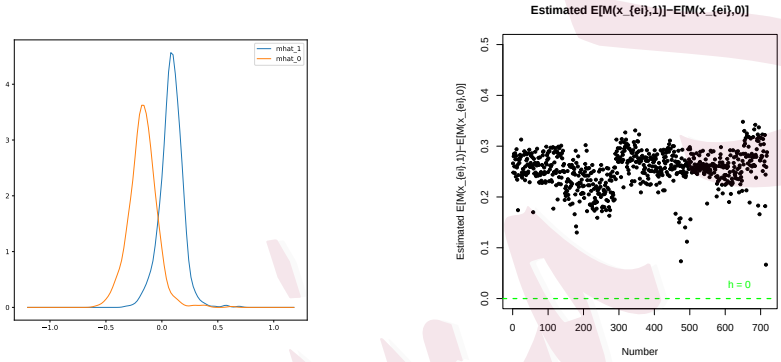

Figure 2: The left panel displays predicted probability density functions of mediators $\mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and $\mathbb{E}[M(\mathbf{x}_{ei}, 1)]$. The right panel illustrates the estimated values of $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ across patient indices.

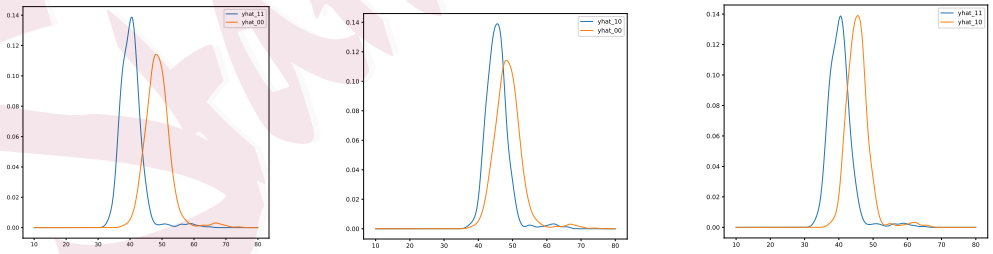

Figure 3: The predicted probability density functions of several $\mathbb{E}[Y(\mathbf{x}_{ei}, t', M_t(\mathbf{x}_{ei}))]$, where $t, t' = 0, 1$.

Using the notations defined in Supplementary Material S7, we first predict $\mathbb{E}[M(\mathbf{x}_{ei}, t)]$ and $\mathbb{E}[Y(\mathbf{x}_{ei}, t', M_t(\mathbf{x}_{ei}))]$ for each $t, t' \in \{0, 1\}$, we denote the pre-

dicted conditional samples as $\{\widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei}), h=1,\ldots,\widehat{n}_e\}$ and $\{\widehat{I}_{\mathbf{Y}}^{(t')}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei}, \widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei})),\ h=1,\ldots,\widehat{n}_e\, j=1,\ldots,\overline{n}_e\}$ for each $t,t'\in\{0,1\}$. Here, we take $\widehat{n}_e=\overline{n}_e=100$. Then, the predicted values are $\widehat{\mathbb{E}}[M(\mathbf{x}_{ei},t)]=\frac{1}{\widehat{n}_e}\left(\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei})\right)$ and $\widehat{\mathbb{E}}[Y(\mathbf{x}_{ei},t',M_t(\mathbf{x}_{ei}))]=\frac{1}{\overline{n}_e\times\widehat{n}_e}\big(\sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(t')}(\overline{\mathbf{z}}_{ej}, \mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(t)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei}))\big)$, for $i=1,2,\cdots,718$.

Figure 2 (left panel) and Figure 3 depict the predicted probability density functions for the predicted values. Based on the prediction result, we can make further discussion below.

7.1 Estimated values of $\mathbb{E}[M(\mathbf{x}_{ei},1)]-\mathbb{E}[M(\mathbf{x}_{ei},0)]$

Given the above predicted values of $\mathbb{E}[M(\mathbf{x}_{ei},t)]$, we can first estimate $\mathbb{E}[M(\mathbf{x}_{ei},1)] -\mathbb{E}[M(\mathbf{x}_{ei},0)]$, which represents the individualized effect of the presence of APOE-ϵ4 alleles on the ventricle volume change. Figure 2 (right panel) shows the predicted results for patients. The predicted density curve of $\mathbb{E}[M(\mathbf{x}_{ei},1)]$ is to the right of the predicted density curve of $\mathbb{E}[M(\mathbf{x}_{ei},0)]$, and all the predicted values of $\mathbb{E}[M(\mathbf{x}_{ei},1)]-\mathbb{E}[M(\mathbf{x}_{ei},0)]$ are positive, suggesting that the presence of APOE-ε4 alleles leads to an expanded proportion of ventricle volume in the whole brain after 12 months. This finding is in accordance with existing medical discoveries (Roussotte et al., 2014) that the presence of

Predictions yielded 718 samples of $\mathbb{E}[M(\mathbf{x}_{ei},0)]$, $\mathbb{E}[M(\mathbf{x}_{ei},1)]$, and $\mathbb{E}[Y(\mathbf{x}_{ei},t',M_t(\mathbf{x}_{ei}))]$ for $t,t'=0,1$, followed by the utilization of the *scipy.stats.gaussian_kde* function from the *SciPy* library in Python to obtain predicted probability density functions.

APOE-ε4 alleles is strongly associated with ventricle expansion.

7.2 Individualized causal effects (ICEs)

Given the above predicted values of $\mathbb{E}[Y(\mathbf{x}_{ei}, t', M_t(\mathbf{x}_{ei}))]$, we estimate three kinds of ICEs, with survival outcome for each of the 718 patients as follows:

$$\widehat{\xi}(1;\mathbf{x}_{ei}) = \frac{1}{\overline{n}_e \times \widehat{n}_e}\left(\sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei})) - \sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei}))\right),$$

$$\widehat{\zeta}(0;\mathbf{x}_{ei}) = \frac{1}{\overline{n}_e \times \widehat{n}_e}\left(\sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei})) - \sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei}))\right),$$

$$\widehat{\tau}(\mathbf{x}_{ei}) = \frac{1}{\overline{n}_e \times \widehat{n}_e}\left(\sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(1)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(1)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei})) - \sum_{j=1}^{\overline{n}_e}\sum_{h=1}^{\widehat{n}_e}\widehat{I}_{\mathbf{Y}}^{(0)}(\overline{\mathbf{z}}_{ej},\mathbf{x}_{ei},\widehat{I}_{\mathbf{M}}^{(0)}(\widehat{\mathbf{z}}_{eh},\mathbf{x}_{ei}))\right),$$

where $i = 1, \ldots, 718$. The alternative decomposition with $\xi(0;\mathbf{x}_{ei})$ and $\zeta(1;\mathbf{x}_{ei})$ can also be used with similar procedures.

Figure 4 presents the estimation results. We can draw two conclusions. First, all values in each subfigure are negative, suggesting that the presence of APOE-ε4 alleles tends to shorten the time to AD onset not only directly but also indirectly by expanding the ventricle. This finding is consistent with the evidence in the literature (Roussotte et al., 2014; Thompson et al., 2004) and is confirmed by Figure 3, where the predicted density curve of $\mathbb{E}[Y(\mathbf{x}_{ei}, 1, M_1(\mathbf{x}_{ei}))]$ is to the left of $\mathbb{E}[Y(\mathbf{x}_{ei}, 0, M_0(\mathbf{x}_{ei}))]$, $\mathbb{E}[Y(\mathbf{x}_{ei}, 1, M_0(\mathbf{x}_{ei}))]$ is to the left of $\mathbb{E}[Y(\mathbf{x}_{ei}, 0, M_0(\mathbf{x}_{ei}))]$, and $\mathbb{E}[Y(\mathbf{x}_{ei}, 1, M_1(\mathbf{x}_{ei}))]$ is to the left of $\mathbb{E}[Y(\mathbf{x}_{ei}, 1, M_0(\mathbf{x}_{ei}))]$. Second, the magnitudes of individual NDE ($\zeta(1;\mathbf{x}_{ei})$) are overall smaller than those of individual NIE ($\xi(0;\mathbf{x}_{ei})$), revealing that the presence of APOE-ε4 alleles contributes to AD onset through the mediated

mechanism more significantly than through the direct path.

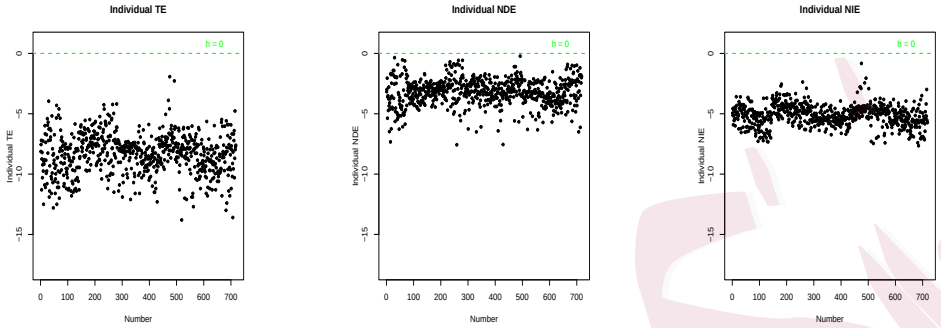

Figure 4: Estimated values of three ICEs in terms of patient index.

We also use the five other methods to estimate $\mathbb{E}[M(\mathbf{x}_{ei}, 1)] - \mathbb{E}[M(\mathbf{x}_{ei}, 0)]$ and the three ICEs defined above. See details in Supplementary Material S10.

7.3 Group average causal effects (GACEs) for subgroups

This section considers covariate-specific groups to examine the relationship between causal effects and these covariates. We define GACEs as coarser than ICEs for $x_1, x_2, \cdots, x_6$ as follows. Abrevaya et al. (2015) introduced conditional average treatment effects (CATEs), and Knaus (2022) and Knaus et al. (2021) further distinguished two cases of CATEs: group average treatment effects and individualized average treatment effects. We now generalize their ideas to define and estimate three kinds of GACEs for $c = 1, 2, \cdots, 6$:

$$\xi_{c,g}(1; \mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, 1, M_1(\mathbf{x})) - Y(\mathbf{x}, 1, M_0(\mathbf{x}))|x_c = g],$$

$$\zeta_{c,g}(0; \mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, 1, M_0(\mathbf{x})) - Y(\mathbf{x}, 0, M_0(\mathbf{x}))|x_c = g],$$

$$\tau_{c,g}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x}, 1, M_1(\mathbf{x})) - Y(\mathbf{x}, 0, M_0(\mathbf{x}))|x_c = g].$$

Among the six covariates $(x_1, x_2, \cdots, x_6)$, x_1 and x_3 are continuous, while the rest are discrete. We follow Knaus (2022) to begin by estimating the above

GACEs along discrete variables, such as gender (x_2), ethnicity (x_4), race (x_5), and marital status (x_6), using the ordinary least squares (OLS) regression. Table 1 presents the results of coefficients and their standard errors. Panel A shows the results of an OLS regression with a male dummy as a covariate, $\tau(\mathbf{x}_{ei})$, $\zeta(0;\mathbf{x}_{ei})$, and $\xi(1;\mathbf{x}_{ei}) = \beta_0 + \beta_1 male_i + error_i$, where β_0 represents the GACE value for the female group, and β_1 indicates how much the GACE differs for the male group. Panel B replaces the male dummy in the regression with a Hispanic or Latino dummy, and similarly for Panel C and D. As shown in Table 1, all coefficients are significant for Panel A, indicating gender difference; the GACEs of males are less than those of females. For example, the group average TE ($\tau_{2,g}(\mathbf{x})$) of males is -8.805 ($-7.688 - 1.117$), revealing that the APOE-ε4–AD association is weaker among females. In contrast, the coefficient for the group average TE ($\tau_{4,g}(\mathbf{x})$) is insignificant for Panel B, implying no notable heterogeneity in GACEs across ethnicity. Panel C shows some racial differences. First, the coefficient of the group average TE ($\tau_{5,g}(\mathbf{x})$) and the group average NIE ($\xi_{5,g}(1;\mathbf{x})$) are significant, suggesting that the effect heterogeneity related to race is manifested mainly by the indirect mechanism. Second, the magnitudes of GACEs in the white group are greater than those of the non-white group, revealing that the APOE-ε4 –AD association is stronger among the white group, aligning with existing medical

findings (Tang et al., 1998). Finally, Panel D shows that the effect heterogeneity related to marital status manifested mainly by the direct mechanism, with a stronger APOE-ε4–AD association among married individuals.

Table 1: Coefficients and standard errors (in parentheses) of GACEs

Panel		$\tau_{c,g}(\mathbf{x})$	$\zeta_{c,g}(0;\mathbf{x})$	$\xi_{c,g}(1;\mathbf{x})$
A	Constant	-7.688^{**} (0.096)	-2.672^{**} (0.062)	-5.016^{**} (0.053)
	Male	-1.117^{**} (0.121)	-0.924^{**} (0.077)	-0.192^{**} (0.067)
B	Constant	-8.326^{**} (0.063)	-3.157^{**} (0.039)	-5.169^{**} (0.031)
	Hispanic or Latino	-0.215 (0.543)	-1.832^{**} (0.413)	1.613^{**} (0.246)
C	Constant	-7.649^{**} (0.263)	-3.057^{**} (0.195)	-4.593^{**}(0.135)
	White	-0.726^{**} (0.271)	-0.158 (0.199)	-0.568^{**} (0.139)
D	Constant	-7.279^{**} (0.119)	-2.372^{**} (0.081)	-4.908^{**} (0.066)
	Married	-1.352^{**} (0.137)	-1.071^{**} (0.091)	-0.281 (0.075)

Note: * and ** correspond to $p < 0.05$ and $p < 0.01$, respectively.

7.4 Nonparametric GACEs for continuous covariates

For the two continuous covariates, age (x_1) and education level (x_3), we use kernel regression (Knaus, 2022) based on the R-package np (Hayfield and Racine, 2008) to estimate GACEs. Figures 5 and 6 present the results.

Figure 5 shows that all three kinds of GACEs are associated with age, and they all slightly increase with age, revealing that the incidence of AD caused by APOE-ε4 alleles decreases with age. This result partly agrees with the existing medical finding (Blacker et al., 1997) that the risk conferred by APOE-ε4 was most marked in the 61 to 65 age group when the sample was stratified on family mean age at onset. However, as shown in Figure 6, all

three kinds of GACEs are associated with education level, and they decrease at first and then remain unchanged with the education level. These results reveal that the impact of APOE-ε4 alleles on the incidence of AD increases initially, then remains unchanged with the education level.

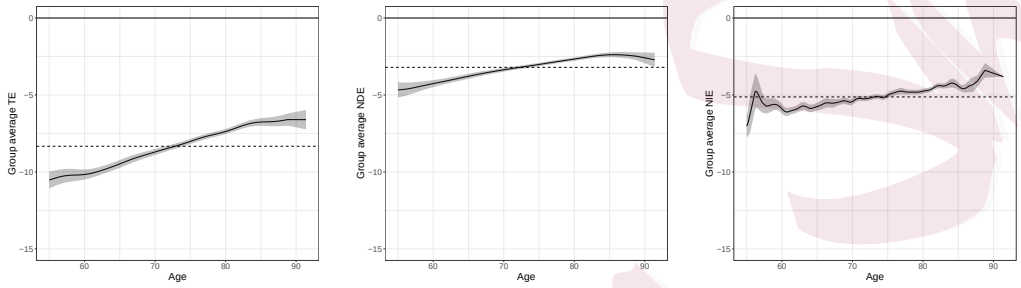

Figure 5: Effect heterogeneity for age. "Group average TE" means estimate of $\tau_{1,g}(\mathbf{x})$ and so on. Dotted lines indicate estimates of the average causal effects and grey areas show the 95%-confidence intervals.

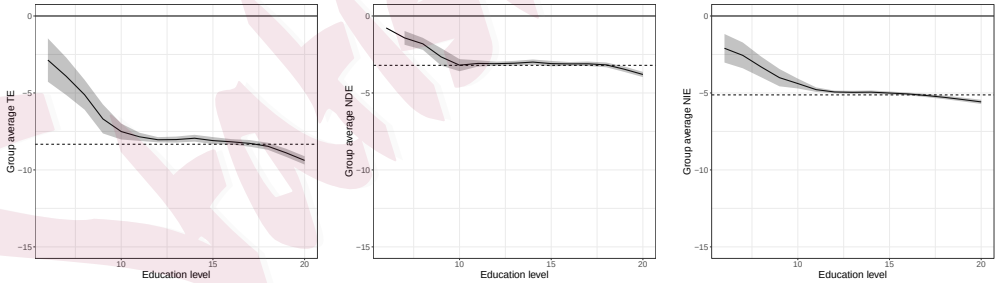

Figure 6: Effect heterogeneity regarding education level, where "Group average TE" means estimate of $\tau_{3,g}(\mathbf{x})$ and so on. Dotted lines indicate estimates of the average causal effects and grey areas show 95%-confidence intervals.

We also model GACEs using the multivariate OLS regression with six covariates and obtain the average causal effects based on ICEs. Detailed

results are presented in Supplementary Material S10.

8. Discussion

This study introduces CGAN-ICMA-SO, an innovative approach for estimating ICEs and exploring individualized causal mechanisms with survival outcomes. Our model is built on the foundation of the CGAN framework and our theoretical backing for CGAN-ICMA-SO is robust, as we underline the convergence of the estimated distribution from our inferential conditional generator to the true conditional distribution. This convergence is assured under certain mild conditions. Through simulation studies, we demonstrate that CGAN-ICMA-SO outperforms five other cutting-edge methods, as measured by the proposed metrics. We further employ CGAN-ICMA-SO to estimate the ICEs of APOE-$\varepsilon 4$ alleles on the time to AD and further investigate the variation of these causal effects in relation to observable characteristics.

Several areas warrant future exploration. From a theoretical standpoint, it would be beneficial to derive the convergence rate of the sampling distribution, thereby fortifying our results. Moreover, future research could productively extend our approach to accommodate more varied types of treatments, such as categorical and continuous treatments. Finally, adapting our method to accommodate multiple-mediator scenarios could significantly broaden its applicability to complex causal mediation analysis.

While our study highlights the strengths and potential of CGAN-ICMA-SO, it is essential to address its limitations. First, the complexity of the model can make its interpretability challenging. Understanding its inner workings could be non-trivial tasks. Second, training the model is computationally demanding, and scaling it to larger datasets or more complex scenarios requires significant resources. Finally, sensitivity analysis is crucial for assessing the robustness of causal conclusions when the unconfoundedness assumption may be violated. While methods like the E-value (VanderWeele and Ding, 2017) quantify the minimum confounder strength required to explain an observed treatment-outcome association, they cannot handle treatment-induced confounding in mediation analysis. Our study lacks such sensitivity analysis for individualized mediation methods. Established strategies tailored for parametric mediation models focusing on average mediation causal effects encompass several key approaches. This involves examining sensitivity parameters such as the error correlation between the mediator and outcome models and the proportion of unexplained variance in the outcome explained by introducing treatment-mediator interaction terms (Imai and Yamamoto, 2013). Another strategy models the joint distribution of potential mediators and outcomes, along with expected value of potential outcomes, using a Gaussian copula model (Albert and Wang, 2015). Additionally, a method involves in-

corporating a latent binary variable U, which denotes the presence or absence of an unmeasured confounder, into the exposure-mediator, exposure-outcome, and mediator-outcome relationships simultaneously. It then compares the estimated causal effects with those derived by disregarding the existence of U under varying prior beliefs on the U-related coefficients (Zhou and Song, 2021). While recent methods have targeted ITE or CATE, Jesson et al. (2021) presents a new parametric interval estimator for high-dimensional datasets. This estimator can determine a range of possible CATE values when provided with a predefined bound for hidden confounding. Jin et al. (2023) proposes a model-free framework of ITE, building upon ideas from conformal inference. Their approach calculates a Γ-value, representing the minimum strength of confounding necessary to invalidate the evidence for ITE. Oprescu et al. (2023) introduces the B-Learner, a meta-learner capable of efficiently establishing sharp bounds on the CATE function within specified constraints on hidden confounding. Yin et al. (2024) proposes a marginal sensitivity model and adapts conformal inference principles to estimate an ITE interval at a given confounding strength. These methodologies predominantly address ITE or CATE scenarios. Expanding these approaches to our individualized mediation analysis method presents a promising direction for future research, necessitating further exploration and in-depth investigation.

Acknowledgements

This research is fully supported by GRF grant (No. 14303622) of HKSAR.

References

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics 33*(4), 485–505.

Albert, J. M. and W. Wang (2015). Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics 16*(2), 339–351.

Bica, I., J. Jordon, and M. van der Schaar (2020). Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems 33*, 16434–16445.

Blacker, D., J. Haines, L. Rodes, H. Terwedow, R. Go, L. Harrell, R. Perry, S. Bassett, G. Chase, D. Meyers, et al. (1997). Apoe-4 and age at onset of alzheimer's disease: the nimh genetics initiative. *Neurology 48*(1), 139–147.

Chapfuwa, P., S. Assaad, S. Zeng, M. J. Pencina, L. Carin, and R. Henao (2021). Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145.

Chapfuwa, P., C. Tao, C. Li, C. Page, B. Goldstein, L. C. Duke, and R. Henao (2018). Adversarial time-to-event modeling. In *International Conference on Machine Learning*, pp. 735–744. PMLR.

REFERENCES

Chen, P., W. Dong, X. Lu, U. Kaymak, K. He, and Z. Huang (2019). Deep representation learning for individualized treatment effect estimation using electronic health records. *Journal of Biomedical Informatics 100*, 103303.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian Additive Regression Trees.

Chu, J., W. Dong, J. Wang, K. He, and Z. Huang (2020). Treatment effect prediction with adversarial deep learning using electronic health records. *BMC Medical Informatics and Decision Making 20*(4), 1–14.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Cui, Y., M. R. Kosorok, E. Sverdrup, S. Wager, and R. Zhu (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(2), 179–211.

Curth, A., C. Lee, and M. van der Schaar (2021). Survite: Learning heterogeneous treatment effects from time-to-event data. *Advances in Neural Information Processing Systems 34*, 26740–26753.

Dyachenko, T. L. and G. M. Allenby (2018). Bayesian analysis of heterogeneous mediation. *Georgetown McDonough School of Business Research Paper* (2600140).

Ge, Q., X. Huang, S. Fang, S. Guo, Y. Liu, W. Lin, and M. Xiong (2020). Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in Genetics 11*, 585804.

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research 50*(1), 1–22.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*, 1–32.

Henderson, N. C., T. A. Louis, G. L. Rosner, and R. Varadhan (2020). Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics 21*(1), 50–68.

Huan, C., X. Song, and H. Yuan (2024). Individualized causal mediation analysis with continuous treatment using conditional generative adversarial networks. *Statistics and Computing 34*(5), 170.

Huang, Y.-T. and H.-I. Yang (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.) 28*(3), 370.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309.

Imai, K. and T. Yamamoto (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis 21*(2), 141–171.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random Survival Forests.

Jesson, A., S. Mindermann, Y. Gal, and U. Shalit (2021). Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pp. 4829–4838. PMLR.

Jin, Y., Z. Ren, and E. J. Candès (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences 120*(6), e2214889120.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal 25*(3), 602–627.

Knaus, M. C., M. Lechner, and A. Strittmatter (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal 24*(1), 134–161.

Martinussen, T., S. Vansteelandt, M. Gerster, and J. v. B. Hjelmborg (2011). Estimation of direct effects for survival data by using the aalen additive hazards model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 773–788.

Oprescu, M., J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, and U. Shalit (2023). B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, pp. 26599–26618. PMLR.

Park, S. and D. Kaplan (2015). Bayesian causal mediation analysis for group randomized designs with homogeneous and heterogeneous effects: Simulation and case study. *Multivariate Behavioral Research 50*(3), 316–333.

Preacher, K. J., D. D. Rucker, and A. F. Hayes (2007). Addressing moderated mediation hypothe-

ses: Theory, methods, and prescriptions. *Multivariate Behavioral Research 42*(1), 185–227.

Qin, X. and G. Hong (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics 42*(3), 308–340.

Qin, X. and L. Wang (2023). Causal moderated mediation analysis: Methods and software. *Behavior Research Methods*, 1–21.

Roussotte, F. F., B. A. Gutman, S. K. Madsen, J. B. Colby, K. L. Narr, P. M. Thompson, A. D. N. I. (ADNI, et al. (2014). The apolipoprotein e epsilon 4 allele is associated with ventricular expansion rate and surface morphology in dementia and normal aging. *Neurobiology of Aging 35*(6), 1309–1317.

Royston, P. and M. K. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine 30*(19), 2409–2421.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association 100*(469), 322–331.

Shen, J., L. Wang, S. Daignault, D. E. Spratt, T. M. Morgan, and J. M. Taylor (2018). Estimating the optimal personalized treatment strategy based on selected variables to prolong survival via random survival forest with weighted bootstrap. *Journal of Biopharmaceutical Statistics 28*(2), 362–381.

Sparapani, R. A., B. R. Logan, R. E. McCulloch, and P. W. Laud (2016). Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in Medicine 35*(16), 2741–2753.

Sun, R., X. Zhou, and X. Song (2021). Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling: A Multidisciplinary Journal 28*(5), 778–790.

Tang, M.-X., Y. Stern, K. Marder, K. Bell, B. Gurland, R. Lantigua, H. Andrews, L. Feng, B. Tycko, and R. Mayeux (1998). The apoe-ϵ4 allele and the risk of alzheimer disease among african americans, whites, and hispanics. *Jama 279*(10), 751–755.

Thompson, P. M., K. M. Hayashi, G. I. De Zubicaray, A. L. Janke, S. E. Rose, J. Semple, M. S. Hong, D. H. Herman, D. Gravano, D. M. Doddrell, et al. (2004). Mapping hippocampal and ventricular change in alzheimer disease. *Neuroimage 22*(4), 1754–1766.

VanderWeele, T. and S. Vansteelandt (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods 2*(1), 95–115.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.) 22*(4), 582.

VanderWeele, T. J. and P. Ding (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine 167*(4), 268–274.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine 11*(14-15), 1871–1879.

Xu, S., L. Liu, and Z. Liu (2022). Deepmed: Semiparametric causal mediation analysis with debiased deep learning. *Advances in Neural Information Processing Systems 35*, 28238–28251.

Xue, F., X. Tang, G. Kim, K. C. Koenen, C. L. Martin, S. Galea, D. Wildman, M. Uddin, and

REFERENCES

A. Qu (2022). Heterogeneous mediation analysis on epigenomic ptsd and traumatic stress in a predominantly african american cohort. *Journal of the American Statistical Association* (just-accepted), 1–36.

Yin, M., C. Shi, Y. Wang, and D. M. Blei (2024). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association 119*(545), 122–135.

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. and X. Song (2021). Mediation analysis for mixture cox proportional hazards cure models. *Statistical Methods in Medical Research 30*(6), 1554–1572.

Cheng Huan

Department of Statistics, The Chinese University of Hong Kong, Hong Kong

E-mail: huancheng@link.cuhk.edu.hk.

Xinyuan Song

Department of Statistics, The Chinese University of Hong Kong, Hong Kong

E-mail: xysong@cuhk.edu.hk.

Hongwei Yuan

Department of Mathematics, University of Macau, Taipa, Macau, China.

E-mail: hwyuan@um.edu.mo.