Counterfactual Estimation via Temporal-Aware Intervention Networks

Xin Wang[1] , **Chi Luo**[1] , **Shengfei Lyu**[2] , **Yi Wan**[1] , **Xiren Zhou**[1] , **Xiangyu Wang**[1] , **Huanhuan Chen**[1,*]

[1]University of Science and Technology of China
[2]Nanyang Technological University
wxin0126@ustc.edu.cn, rogan010603@gmail.com, shengfei.lyu@ntu.edu.sg, yi_wan@mail.ustc.edu.cn, {zhou0612, sa312, hchen}@ustc.edu.cn

Abstract

Accurate estimation of time-varying treatment effects is crucial for optimizing interventions in personalized medicine. However, observational data often contains complex confounding bias and temporal complexities, making counterfactual estimation challenging. We propose Counterfactual Estimation via Temporal-Aware Intervention Networks (TAIN), a novel model that introduces an Intervention-aware Functional Convolution kernel to emphasize the role of treatments and capture complex temporal treatment interactions. TAIN addresses confounding bias from a domain generalization perspective, approximating the unknown target domain using adversarial examples and incorporating Sharpness-Aware Minimization to derive a generalization bound. This approach is more suitable for longitudinal settings compared to existing methods inspired by domain adaptation techniques due to inherent differences between static and longitudinal contexts. Experiments on simulated datasets demonstrate TAIN's superior performance compared to state-of-the-art models for counterfactual estimation over time.

1 Introduction

Precise estimation of time-varying treatment effects is crucial in domains like personalized medicine, where optimizing individual interventions relies on accurate estimating counterfactual outcomes [Huang and Ning, 2012]. Randomized Controlled Trials, despite being the gold standard for causal inference [Hariton and Locascio, 2018], are limited by high costs and ethical concerns. Consequently, research has shifted focus to advanced methods tackling confounding bias and temporal complexities in observational data for accurate counterfactual estimation.

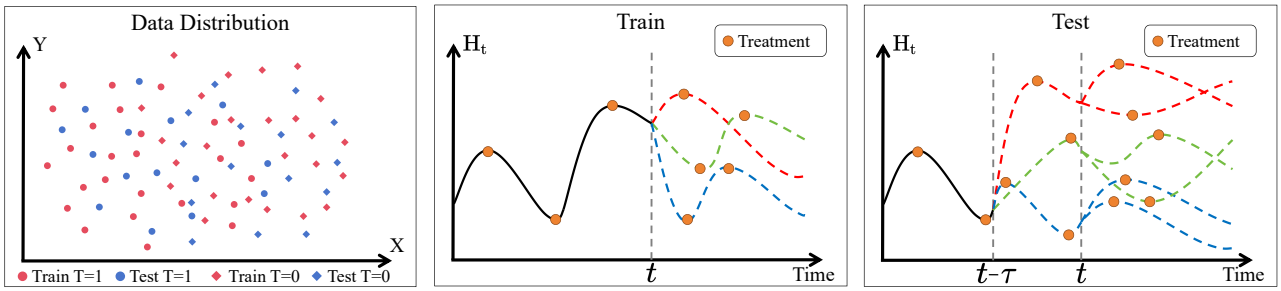
Time-varying confounders often introduce complex confounding bias in observational data, leading to inaccurate estimates. Recent studies [Bica *et al.*, 2020b; Melnychuk *et al.*, 2022; Wang *et al.*, 2024] have addressed this issue by learning representations to break the association between historical information and treatment assignment, drawing inspiration from domain adaptation techniques used in static causal inference settings [Shalit *et al.*, 2017; Hassanpour and Greiner, 2019; Johansson *et al.*, 2022]. However, domain adaptation's effectiveness in longitudinal settings remains questionable due to inherent differences between the contexts.

Due to the unobservability of counterfactuals, we aim to train a model on the factual (source) domain that generalizes well to the counterfactual (target) domain. In static settings (Figure 1(a)), the target domain during testing is consistent with training and can be sampled, aligning with domain adaptation [Ganin *et al.*, 2016]. However, in longitudinal settings, the historical information of length t generated under the training intervention policy (Figure 1(b)) differs from the τ-step intervened historical information under a different testing policy (Figure 1(c)). The unknown data generation process makes the target distribution unobservable during testing, possibly explaining why current balancing methods fail in longitudinal settings [Huang *et al.*, 2024].

Another key challenge in longitudinal settings lies in modeling temporal treatment interactions. Take antibiotic treatments as an example - [Roemhild *et al.*, 2022] have shown that the effectiveness of current antibiotics strongly depends on previously administered ones, making it critical to understand these temporal interactions for optimal treatment. While recent work ACTIN [Wang *et al.*, 2024] attempts to capture such interactions using a dual-module architecture, it follows previous approaches [Bica *et al.*, 2020b; Melnychuk *et al.*, 2022] in treating treatments merely as regular inputs. This fails to emphasize treatments' special role in determining outcomes. Though the importance of emphasizing treatments has been demonstrated in static settings by VCNet [Nie *et al.*, 2020], how to properly model treatment interactions over time remains an open challenge.

To address these challenges, we propose a novel model called Temporal-Aware Intervention Networks (TAIN) for counterfactual estimation over time. In order to emphasize the role that treatments play, we draw inspiration from VCNet, which allows the weights of the prediction head to be functions of the treatment. TAIN introduces an Intervention-aware [1] Functional Convolution (IFC) kernel, ensuring that

[1]In causal inference, "intervention" is a broad term. We primarily use "treatment" for medical interventions, although the terms are used interchangeably herein.



(a) Static setting (b) Traing in longitudinal setting (c) Testing in longitudinal setting

Figure 1: In static settings, the allocation of treatments does not affect the generation of covariates. As illustrated in Figure 1(a), confounding bias leads to distributional differences among intervention groups, while the distribution within each treatment group remains consistent across training and testing sets. However, in longitudinal settings, the data distribution is determined by a sequence of treatments. When the intervention policy during the testing phase from time $t-\tau$ to t (Figure 1(c)) differs from that of the training phase (Figure 1(b)), it is infeasible to sample from the true distribution, as the data generation mechanism is unknown.

the convolutional kernel at each time step depends on the related treatment. Specifically, we employ Radial Basis Functions (RBFs) as basis functions to fit nonlinear functions of treatments, enhancing the model's ability to handle complex temporal treatment interactions. Furthermore, since the target domain is unobservable in longitudinal settings, we revisit confounding bias from the perspective of domain generalization. TAIN approximates the unknown target domain by generating adversarial examples using the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015]. By incorporating Sharpness-Aware Minimization (SAM) [Foret *et al.*, 2021], a technique for enhancing model generalization performance, we derive a generalization bound for the model under certain conditions. The objective function proposed based on this theoretical analysis can more effectively mitigate time-varying confounding bias. Our main contributions are summarized as follows:

- We propose a novel Intervention-aware Functional Convolution (IFC) kernel to capture complex temporal treatment interactions by designing convolutional kernels as functions of treatments.
- We revisit time-varying confounding bias from domain generalization perspective and derive a generalization bound using adversarial examples with SAM.
- Extensive experiments on synthetic datasets demonstrate TAIN's superior performance in counterfactual estimation over time.

2 Problem Formulation

Consider an i.i.d. observational dataset $\mathcal{D}$ containing detailed information for N patients, denoted as $\mathcal{D} = \left\{ \mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}, \mathbf{y}_t^{(i)}{}_{t=1}^{T^{(i)}} \cup \mathbf{v}^{(i)} \right\}_{i=1}^{N}$. For each patient i, we observe time-varying covariates $\mathbf{X}_t^{(i)} \in \mathcal{X}$, treatments $\mathbf{A}_t^{(i)} \in \mathcal{A}$, and outcomes $\mathbf{Y}_t^{(i)} \in \mathcal{Y}$ at discrete time steps $T^{(i)}$, along with static covariates $\mathbf{V}^{(i)} \in \mathcal{V}$ such as gender and age. For notational simplicity, we omit the patient-specific superscript (i) when the context is clear.

Building upon the potential outcome framework [Rubin, 1978] and its extension to time-varying treatments [Robins and Hernan, 2008], we aim to estimate time-varying counterfactual outcomes [Bica *et al.*, 2020b; Lim *et al.*, 2018; Li *et al.*, 2021]. Let $\bar{\mathbf{H}}_t = (\bar{\mathbf{X}}_t, \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{Y}}_t, \mathbf{V})$ denote the patient's historical information, where $\bar{\mathbf{X}}_t = (\mathbf{X}_1, \cdots, \mathbf{X}_t)$, $\bar{\mathbf{Y}}_t = (\mathbf{Y}_1, \cdots, \mathbf{Y}_t)$, and $\bar{\mathbf{A}}_{t-1} = (\mathbf{A}_1, \cdots, \mathbf{A}_{t-1})$. Our objective is to estimate the potential outcome $\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}]$ following a treatment sequence $\bar{\mathbf{a}}_{t:t+\tau-1} = (\mathbf{a}_t, \cdots, \mathbf{a}_{t+\tau-1})$, conditioned on the historical information $\bar{\mathbf{H}}_t$:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}] | \bar{\mathbf{H}}_t]. \tag{1}$$

To identify treatment effects from observational data, we rely on the assumptions of consistency, sequential ignorability, and sequential overlap, as established in prior works [Lim *et al.*, 2018; Bica *et al.*, 2020b; Melnychuk *et al.*, 2022; Wang *et al.*, 2024] and detailed in Appendix B.

3 Related Work

3.1 Causal inference with longitudinal data

Recently, significant progress has been made in the field of longitudinal data analysis by incorporating neural network techniques for counterfactual estimation. For instance, RMSN [Lim *et al.*, 2018] combines two propensity networks and employs an inverse probability weighting (IPTW) based training method in its prediction model. G-Net [Li *et al.*, 2021] enhances the traditional G-computation technique through a deep learning framework. Similarly, CRN [Bica *et al.*, 2020b], which is also based on RNN networks, focuses on learning balanced representations to mitigate confounding bias. In contrast to these works, CT [Melnychuk *et al.*, 2022] proposes utilizing a more powerful transformer model to better handle long-term dependencies, while ACTIN [Wang *et al.*, 2024] introduces a dual-module framework that effectively enhances the ability of simple models to handle complex temporal treatment interactions. It is worth noting that CRN, CT, and ACTIN all adopt the idea of domain adaptation to learn balanced representations by eliminating the association between historical information and current treatment assignments to alleviate confounding bias.

There have been significant advances in causal inference based on longitudinal data, albeit with different settings from ours [Bica *et al.*, 2020a; Frauen *et al.*, 2023; Seedat *et al.*, 2022; Meng *et al.*, 2023; Hess *et al.*, 2024]. These works primarily focus on addressing time-varying confounding factors in longitudinal observational data. For instance, the Time Series Deconfounder [Bica *et al.*, 2020a] and DeepACE [Frauen *et al.*, 2023] utilize deep learning models to adjust for time-varying confounders, while TE-CDE [Seedat *et al.*, 2022] and BNCDE [Hess *et al.*, 2024] propose differential equation-based approaches for irregularly sampled data. A detailed discussion of these related works is provided in Appendix A.

3.2 Domain generalization

Domain generalization (DG) aims to learn a model that can perform well on unseen target domains, which is crucial for addressing distribution shift problems in real-world scenarios [Muandet *et al.*, 2013]. Many methods have been proposed to tackle this issue, including learning domain-invariant representations [Motiian *et al.*, 2017; Matsuura and Harada, 2020], data augmentation-based approaches [Xu *et al.*, 2020], and robustness training-based methods [Shi *et al.*, 2021; Foret *et al.*, 2021; Zhang *et al.*, 2024]. Despite the significant progress made by these methods in tasks such as image classification, their application to time series data remains challenging. Compared to static data, the dynamic generative process of time series data makes their distribution more complex, which may be the reason why the strategy of learning domain-invariant representations is less effective in mitigating time-varying confounding bias. The flatness-related methods, such as Sharpness-Aware Minimization (SAM) [Foret *et al.*, 2021], adopted in this paper aim to address the effects of domain shifts by identifying flat minima, seeking regions in the loss landscape where small perturbations in the input have minimal impact on the model's predictions. By leveraging flat minima, these methods enhance the model's robustness to domain variations. This approach represents one possible solution path, offering a domain generalization perspective for mitigating time-varying confounding bias.

4 Method

4.1 Intervention-aware Functional Convolution

To enhance the model's ability to emphasize the role of treatments and capture potentially complex temporal interactions, we propose an Intervention-aware Functional Convolution (IFC) kernel. Explicitly encoding treatment information within the functional convolution operation enables IFC to better understand treatment impact on patient outcomes over time. This intervention-aware approach facilitates the learning of intricate relationships between treatments and covariates, ultimately improving the model's capacity for predicting counterfactual outcomes accurately.

In our proposed IFC, we employ a one-dimensional dilated convolution kernel that differs from traditional approaches [Oord *et al.*, 2016] by incorporating the corresponding treatment as a function at each time step. Specifically, the output at time step t can be computed as:

$$F(t) = \sum_{i=0}^{k-1} \mathbf{W}_i \phi(\mathbf{a}_t) \mathbf{z}_{t-d\cdot i} + \mathbf{b}. \tag{2}$$

In the equation, $F(t) \in \mathbb{R}^{d_{\text{out}}}$ denotes the output at time step t, where d_{out} is the output dimension. The convolution kernel size is denoted by k, while d denotes the dilation factor. $\mathbf{W}_i \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the i-th convolution weight matrix, where d_{in} is the input dimension. The intervention function $\phi(\mathbf{a}_t)$ takes the treatment $\mathbf{a}_t$ at time step t as input. $\mathbf{z}_{t-i} \in \mathbb{R}^{d_{\text{in}}}$ is the input at time step $t-i$, and $\mathbf{b} \in \mathbb{R}^{d_{\text{out}}}$ denotes the bias.

Radial Basis Functions (RBFs) are widely used as basis functions for approximating nonlinear functions and capturing complex interactions between variables. In this study, we explore two types of RBFs, Gaussian and Multiquadric, as the basis functions in the IFC kernel to capture the temporal treatment interactions.

Let $\mathbf{c} \in \mathbb{R}^{d_c}$ denote the centers of the RBFs. The Gaussian RBF for the i-th dimension of the treatment $\mathbf{a}$ and the j-th center $\mathbf{c}$ is defined as:

$$\psi_{\text{Gaussian}}(\mathbf{a}^i, \mathbf{c}^j) = \exp\left(-\frac{(\mathbf{a}^i - \mathbf{c}^j)^2}{2\zeta^2}\right), \tag{3}$$

where ζ is the width parameter controlling the spread of the Gaussian function. Similarly, the Multiquadric RBF for the i-th dimension of $\mathbf{a}$ and the j-th center $\mathbf{c}$ is:

$$\psi_{\text{Multiquadric}}(\mathbf{a}^i, \mathbf{c}^j) = \sqrt{(\mathbf{a}^i - \mathbf{c}^j)^2 + \sigma^2}, \tag{4}$$

where σ is a constant parameter controlling the shape of the Multiquadric function.

The overall basis function $\phi(\mathbf{a})$ is obtained by summing the weighted RBF values over all dimensions of the treatment and all centers:

$$\phi(\mathbf{a}) = \sum_{i=1}^{d_a} \sum_{j=1}^{d_c} w_{ij} \psi(\mathbf{a}^i, \mathbf{c}^j), \tag{5}$$

where $\psi(\mathbf{a}^i, \mathbf{c}^j)$ can be either the Gaussian RBF or the Multiquadric RBF, and $\{w_{ij}\}$ are the learnable weights associated with each RBF. By incorporating these RBF-based basis functions into the IFC, the model can learn the complex, nonlinear relationships in the temporal dynamics of treatment effects, enabling more accurate estimation of counterfactual outcomes.

4.2 Addressing Confounding Bias from a Domain Generalization Perspective

Existing methods for causal inference with time-varying treatments often borrow ideas from the static setting and aim to address confounding bias through the lens of domain adaptation. Specifically, these approaches seek to learn representations that align the conditional distributions $P(\Phi(\bar{\mathbf{H}}_t)|\mathbf{A}_t = \mathbf{a}_j)$ across different treatment assignments $\mathbf{a}_j$, thereby removing the association between time-dependent confounders present in the patient history $\bar{\mathbf{H}}_t$ and time-varying treatments $\mathbf{A}_t$ [Lim *et al.*, 2018; Bica *et al.*, 2020b; Melnychuk *et al.*, 2022; Wang *et al.*, 2024]. However,

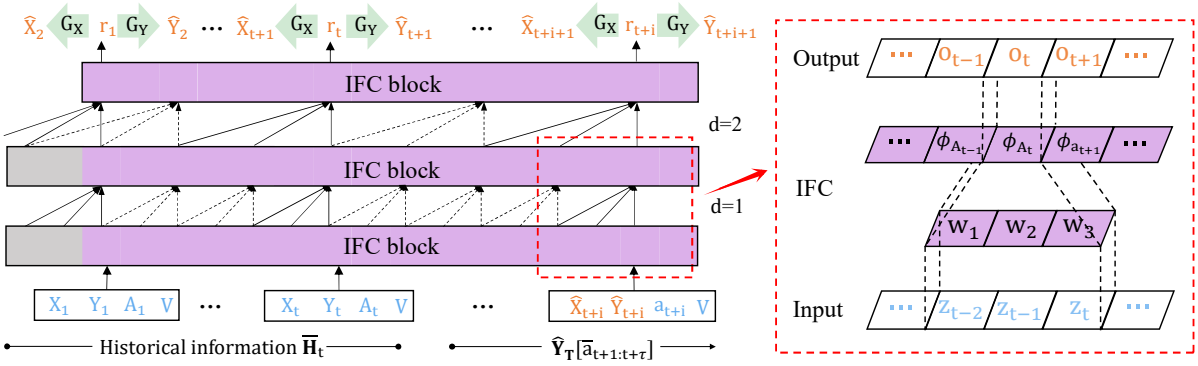


Figure 2: The module in TAIN that learns representations of historical information primarily consists of multiple IFC blocks. To enlarge the receptive field, we employ dilated convolutions, where d denotes the dilation factor. The right side of the figure illustrates the structure of an IFC kernel, which is determined by the treatment nonlinear function $\phi(\mathbf{A}_t)$ at the current time step (abbreviated as $\phi_{\mathbf{A}_t}$ in the figure). Subsequently, counterfactual estimation is performed on the learned representations using feedforward neural networks G_X and G_Y.

for this alignment to successfully remove confounding bias, it is assumed that the distribution of the observed histories $\bar{\mathbf{H}}_t$ remains consistent between training and testing (see Remark C.5 for a detailed explanation). In the longitudinal setting, this assumption may not hold, limiting the effectiveness of these methods in real-world scenarios. To illustrate this, we can derive the distribution of $\bar{\mathbf{H}}_t$ in the training phase as:

$$P(\bar{\mathbf{H}}_t) = P(\mathbf{X}_1) \prod_{s=1}^{t-1} P_{\text{train}}(\mathbf{A}_s|\bar{\mathbf{H}}_s) P(\mathbf{X}_{s+1}|\bar{\mathbf{H}}_s, \mathbf{A}_s) \quad (6)$$

where $P_{\text{train}}(\mathbf{A}_s|\bar{\mathbf{H}}_s)$ denotes the probability of treatment assignment at step s given the history $\bar{\mathbf{H}}_s$, which is typically determined by doctors following specific treatment policies during training.

However, during the testing phase, given a history $\bar{\mathbf{H}}_{t-\tau}$, we may encounter a new sequence of treatments $\bar{\mathbf{A}}_{t-\tau:t-1}$ of length τ, which results in a trajectory of length t. The distribution of this trajectory can be expressed as:

$$P_\tau(\bar{\mathbf{H}}_t) = P(\bar{\mathbf{H}}_{t-\tau}) \prod_{s=t-\tau}^{t-1} P_{\text{test}}(\mathbf{A}_s|\bar{\mathbf{H}}_s) P(\mathbf{X}_{s+1}|\bar{\mathbf{H}}_s, \mathbf{A}_s), \quad (7)$$

where $P_{\text{test}}(\mathbf{A}_s|\bar{\mathbf{H}}_s)$ represents the treatment assignment policy employed during the testing phase (e.g., random assignment), which generally differs from the policy used during training.

The testing phase involves trajectories $\bar{\mathbf{H}}_t$ of length t, formed by applying different treatment subsequences of length τ, whose distribution often differs from that observed during training. This discrepancy between the distributions of $\bar{\mathbf{H}}_t$ in the training and testing phases poses a significant challenge for methods that rely on the assumption of distributional consistency, leading to the inability to maintain consistent conditional distributions $P(\Phi(\bar{\mathbf{H}}_t)|\mathbf{A}_t = \mathbf{a}_j)$ across different treatment assignments $\mathbf{a}_j$ in the testing phase.

In contrast to previous work, this paper proposes a novel perspective on addressing confounding bias through the lens of domain generalization. Let $\mathbf{u}_t = [\bar{\mathbf{H}}_t, \mathbf{A}_t] \in \mathcal{U}$, $h : \mathcal{U} \to \mathcal{Y}$ be a hypothesis function for counterfactual estimation, and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a loss function. We denote the source (observed) domain of $\mathbf{u}_t$ as D_S, which follows the factual distribution $P_S(\mathbf{u}_t) = P(\bar{\mathbf{H}}_t|\mathbf{A}_t)P(\mathbf{A}_t)$, and the target domain as D_T, which follows the counterfactual distribution $P_T(\mathbf{u}_t) = P_{\text{test}}(\bar{\mathbf{H}}_t|\mathbf{A}_t)P_{\text{test}}(\mathbf{A}_t)$. Let $R_S := \mathbb{E}_{\mathbf{u}_t \sim P_S}[l(h(\mathbf{u}_t), \mathbf{y}_t)]$ denote the factual risk of a hypothesis h, and similarly, let R_T denote the counterfactual risk. Our goal is to find a hypothesis h that minimizes R_T.

Previous works [Bica *et al.*, 2020b; Melnychuk *et al.*, 2022; Wang *et al.*, 2024] have often approached this problem from the perspective of domain adaptation. However, based on our earlier analysis, for the target domain D_T, we not only lack knowledge of the potential outcomes but also of its distribution, making it infeasible to sample from the observed data. In other words, we are faced with a domain generalization problem, where the model have to generalize to unseen distributions during testing. To address this challenge, we propose a novel approach that constructs a upper bound on the counterfactual risk based on Sharpness-Aware Minimization (SAM) and adversarially generated samples. The specific details are provided in the next section.

4.3 Generalization Bound on Counterfactual Risk

In order to derive a upper bound on the counterfactual risk R_T, we first introduce the concept of SAM [Foret *et al.*, 2021]. SAM is a novel optimization technique that seeks to find a flat minimum of the loss function, which has been shown to improve the generalization performance of deep learning models. The key idea behind SAM is to minimize the maximum loss within a neighborhood of the current model parameters, rather than just the loss at the current parameters. Formally, given a model h_θ parameterized by θ, SAM solves the following optimization problem:

$$\min_{\theta} \max_{\|\epsilon\| \le \rho} R_S(h_{\theta+\epsilon}), \quad (8)$$

where ϵ is a small perturbation to the model parameters, and ρ is a hyperparameter that controls the size of the neighborhood around θ. By solving this minimax optimization problem, SAM finds a set of parameters θ^* that minimize the maximum loss within a ρ-neighborhood of θ^*, resulting in a flatter loss

landscape and improved generalization. Additionally, [Foret *et al.*, 2021] point out that the factual risk can be bounded as shown in Lemma 4.1.

Lemma 4.1 (Sharpness-Aware Minimization [Foret *et al.*, 2021])**.** *The source risk $R_S(h_\theta)$ is bounded using the following PAC-Bayes generalization bound for any ρ with probability $1-\delta$:*

$$R_S(h_\theta) \leq \max_{\|\epsilon\|\leq\rho} \hat{R}_S(h_{\theta+\epsilon}) + \gamma(\|\theta\|_2^2/\rho^2), \tag{9}$$

where n is the training set size used for calculation of empirical risk $\hat{R}_S(h_\theta)$, k is the number of parameters and $\|\theta\|_2$ is the norm of the weight parameters.

Given that our primary objective is to optimize the counterfactual error, we first introduce the generalization bounds proposed by [Shalit *et al.*, 2017] based on domain adaptation to bridge the gap between the factual and counterfactual risks.

Lemma 4.2 (Generalization Bound via IPM [Shalit *et al.*, 2017])**.** *Let $\Phi : \mathcal{U} \to \mathcal{R}$ be a one-to-one representation function, with inverse Ψ. Let $h : \mathcal{U} \to \mathcal{Y}$ and $f : \mathcal{R} \to \mathcal{Y}$ be hypothesis and H, F be the sets of all possible hypothesis (i.e. Hypothesis Space) over $\mathcal{U}$ and $\mathcal{R}$ respectively. Let $\mathcal{G}$ be a family of functions $g : \mathcal{R} \to \mathcal{Y}$. Define the Integral Probability Metric(IPM) of two distributions:*

$$\mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_T^\Phi) = \sup_{g\in\mathcal{G}} \left| \mathbb{E}_{\mathbf{r}\sim P_S^\Phi}[g(\mathbf{r})] - \mathbb{E}_{\mathbf{r}\sim P_T^\Phi}[g(\mathbf{r})] \right|, \tag{10}$$

where $\mathbf{r} = \Phi(\mathbf{u})$. Suppose there exists a constant $B_\Phi > 0$, such that $\forall h \in H, y \in \mathcal{Y}$, $\frac{1}{B_\Phi} l(f(\mathbf{r}), y) \in G$. Then we have:

$$R_T(h) \leq R_S(h) + B_\Phi \cdot \mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_T^\Phi). \tag{11}$$

However, as analyzed in Section 4.2, we cannot sample from the distribution P_T^Φ, rendering the bound provided in Equation 11 impractical for direct application. To address this issue, we propose to generate adversarial examples using the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015] and provide the following lemma to upper bound $\mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_T^\Phi)$.

Lemma 4.3 (Generalization Bound via Adversarial Distribution)**.** *Let P_S^Φ and P_T^Φ be the source and target distributions of the representation $\mathbf{r}$, respectively. Consider the following conditions:*

1. *For all $f \in F$ and $y \in \mathcal{Y}$, the loss function $l(f(\mathbf{r}), y) \in \mathcal{G}$;*
2. *$\mathcal{G}$ is a set of Lipschitz functions, i.e., there exists $L > 0$ such that for all $g \in \mathcal{G}$ and $\mathbf{r}_1, \mathbf{r}_2 \in \mathcal{R}$, $|g(\mathbf{r}_1) - g(\mathbf{r}_2)| \leq L|\mathbf{r}_1 - \mathbf{r}_2|$;*
3. *$\mathbb{E}_{\mathbf{r}\sim P_A^\Phi}|\mathbf{r}| + \mathbb{E}_{\mathbf{r}\sim P_S^\Phi}|\mathbf{r}| < \infty$;*
4. *$\forall f \in F, \max_y \left| \mathbb{E}_{\mathbf{r}\sim P_A^\Phi}[l(f(\mathbf{r}),y)] - \mathbb{E}_{\mathbf{r}\sim P_S^\Phi}[l(f(\mathbf{r}),y)] \right| > 0$.*

Let the adversarial samples be generated by $\mathbf{r}_{adv} = \mathbf{r} + \epsilon \cdot sign(\nabla_\mathbf{r} l(f(\mathbf{r}), y))$ and $\mathbf{r}_{adv} \sim P_A^\Phi$, where $\epsilon = \arg\max_\epsilon [l(f(\mathbf{r}_{adv}), y)) - l(f(\mathbf{r}), y))]$. If conditions (1)-(4) hold, then there exists a constant $M_\Phi > 0$ such that

$$\mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_T^\Phi) \leq M_\Phi \cdot \mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_A^\Phi). \tag{12}$$

Here, M_Φ is a bounded constant determined by the Lipschitz regularity of $\mathcal{G}$ and the moments of the representation distributions; see Appendix D.

Lemma 4.3 indicates that under certain conditions, we can approximate $\mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_T^\Phi)$ using $\mathrm{IPM}_\mathcal{G}(P_S^\Phi, P_A^\Phi)$, where P_A^Φ denotes the distribution of adversarial examples generated from the source distribution P_S^Φ. Among these conditions, the third one ensures that the adversarial and target distributions are *absolutely integrable*, a property satisfied by probability distributions with finite variance. This condition guarantees that the expected values of functions under these distributions are well-defined and finite. Furthermore, the fourth condition is reasonable because adversarial examples are generated to maximize the model's loss, implying that the loss incurred by the adversarial distribution is likely to be strictly greater than that of the source distribution. This condition ensures that the adversarial distribution provides a meaningful upper bound on the target risk.

Under these conditions, Lemma 4.3 provides a tractable way to approximate the distributional discrepancy between the source and target domains using the IPM between the source distribution and its adversarially perturbed counterpart. This approximation allows us to optimize the representation Φ to minimize the distributional discrepancy, even without having access to samples from the target domain. To further improve the model's generalization ability on unseen domains, we derive an upper bound on the factual risk by combining the results from Lemmas 4.1, 4.2, and 4.3.

Theorem 4.4 (Generalization Bound via SAM and Adversarial Distribution)**.** *Under the conditions of Lemma 4.2 and Lemma 4.3, we have:*

$$R_T(h_\theta) \leq \max_{\|\epsilon\|\leq\rho} \hat{R}_S(h_{\theta+\epsilon}) + \alpha_\Phi \mathrm{IPM}(P_S^\Phi, P_A^\Phi) + \gamma(\|\theta\|_2^2/\rho^2), \tag{13}$$

where $\alpha_\Phi = B_\Phi M_\Phi > 0$ is a theoretical constant.

In the following, we introduce how to leverage the IFC kernel to construct models for counterfactual estimation over time, and optimize them based on theoretical results.

4.4 Counterfactual Estimation

As shown in Figure 2, we propose the Temporal Adaptive Convolutional Intervention Network (TAIN). Following the methodology of [Bai *et al.*, 2018], TAIN consists of a representation module $\Phi(\cdot)$, which employs residual connections [He *et al.*, 2016] to concatenate multiple IFC blocks:

$$\mathbf{o} = \mathrm{Activation}(B(\mathbf{z}) + F(\mathbf{z})). \tag{14}$$

When input $\mathbf{z}$ and output $F(\mathbf{z})$ dimensions match, B is an identity map. Otherwise, B becomes a 1×1 convolution to align dimensions.

TAIN employs an autoregressive recursive strategy [Chevillon, 2007; Taieb and Atiya, 2015] for multi-step prediction, which has also been adopted by [Li *et al.*, 2021; Wang *et al.*, 2024]. During the training process, we need to predict the output and time-varying covariates at the next time step. TAIN defines a feedforward neural network G_Y to decode the predicted output from the representation $\Phi(\mathbf{u}_t)$.

For notational simplicity, we denote $\Phi(\mathbf{u}_t)$ as $\mathbf{r}_t$. We use the Mean Squared Error (MSE) to define the following loss function:

$$\mathcal{L}_Y^t(\theta) = \|\mathbf{Y}_{t+1} - G_Y(\mathbf{r}_t)\|^2. \quad (15)$$

TAIN employs two feedforward neural networks, G_X and J_X, for covariate prediction. Specifically, G_X is responsible for decoding the expected covariates from $\mathbf{r}_t$. Drawing inspiration from [Wang *et al.*, 2024], we utilize J_X to design a smoothing mechanism. This mechanism, influenced by the gating mechanism in GRUs [Cho *et al.*, 2014], aims to adapt to the varying trends of different covariates. Analogously, we define the following loss function:

$$\mathcal{L}_X^t(\theta) = \|\mathbf{X}_{t+1} - (\eta G_X(\mathbf{r}_t) + (1-\eta)\mathbf{X}_t)\|^2, \quad (16)$$

where $\eta = \text{Sigmoid}(J_X(\mathbf{r}_t))$ is the smoothing factor. Following G-formula [Robins, 1986], we estimate the expected outcomes as:

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}]|\bar{\mathbf{H}}_t] = \sum_{\bar{\mathbf{l}}_{t:t+\tau-1}} \mathbb{E}[\mathbf{Y}_{t+\tau}|\bar{\mathbf{a}}_{t:t+\tau-1}, \bar{\mathbf{l}}_{t:t+\tau-1}, \bar{\mathbf{H}}_t] \times \prod_{k=t}^{t+\tau-1} f\{\mathbf{l}_k|\bar{\mathbf{a}}_{t:k}, \bar{\mathbf{l}}_{t:k-1}, \bar{\mathbf{H}}_t\}, \quad (17)$$

where $\bar{\mathbf{L}}_t$ represents the history of outputs and covariates up to time t. During the prediction phase, we approximate this expectation by sampling the predicted values from Gaussian distributions with means equal to the model predictions and standard deviation σ. Based on Theorem 4.4, we propose the following objective function:

$$\mathcal{L} = \frac{1}{N}\sum_{i\in\mathcal{D}}\sum_{t=1}^{T}\max_{\|\epsilon\|\leq\rho}(\mathcal{L}_Y^{t(i)}(\theta+\epsilon) + \lambda_1\mathcal{L}_X^{t(i)}(\theta+\epsilon)) + \lambda_2\sum_{t=1}^{T}\text{IPM}(\mathbf{r}_t^{i\in\mathcal{D}}, \mathbf{r}_{t_{adv}}^{i\in\mathcal{D}}) + \lambda_3\mathscr{R}(\theta), \quad (18)$$

where $\mathscr{R}(\theta)$ is the regularization term.

5 Experiments

We compare our method against state-of-the-art models for counterfactual estimation over time: **RMSN** [Lim *et al.*, 2018], **CRN** [Bica *et al.*, 2020b], **G-Net** [Li *et al.*, 2021], **CT** [Melnychuk *et al.*, 2022], **ACTIN** [Wang *et al.*, 2024], and **CTD-NKO** [Wang *et al.*, 2025]. To ensure fair comparison, we perform hyperparameter tuning for all baselines (see details in Appendix F), and report RMSE following the standard protocol [Melnychuk *et al.*, 2022] to measure prediction errors against ground-truth counterfactual outcomes.

5.1 Experiments with FS-Tumor data

Data. The fully-synthetic tumor (FS-Tumor) growth dataset, constructed using a pharmacokinetic-pharmacodynamic model [Geng *et al.*, 2017], provides a realistic simulation of the combined effects of chemotherapy and radiotherapy on lung cancer patients. This dataset has gained significant attention in the research community and has been extensively utilized for evaluating various causal inference methods in studies such as [Lim *et al.*, 2018; Bica *et al.*, 2020b; Melnychuk *et al.*, 2022; Wang *et al.*, 2024]. One notable feature of this advanced bio-mathematical model is the inclusion of a parameter γ, which allows for the control of time-varying confounding in the dataset. By adjusting the value of γ, researchers can simulate scenarios where historical data have varying degrees of influence on treatment allocation. As γ increases, the confounding bias becomes more pronounced, as past information plays a more dominant role in determining the course of treatment.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
RMSN	1.85±0.56	3.03±0.76	3.50±0.83	3.86±0.79	4.59±1.44	5.21±1.38
CRN	1.67±0.39	2.56±0.78	3.07±0.90	3.52±0.89	4.19±1.53	4.77±1.48
G-Net	2.06±0.54	2.92±0.84	3.48±1.03	3.80±1.01	4.42±1.48	5.11±1.52
CT	1.85±0.49	2.81±0.83	3.30±0.87	3.63±0.87	4.20±1.35	4.85±1.37
ACTIN	0.89±0.31	1.48±0.61	1.94±0.69	2.58±0.84	3.31±1.22	4.24±1.83
CTD-NKO	0.76±0.53	1.70±0.87	2.28±1.16	2.72±1.26	3.18±1.59	3.95±1.85
TAIN	**0.39±0.10**	**0.60±0.16**	**0.82±0.30**	**1.13±0.40**	**1.31±0.48**	**1.63±0.58**

Table 1: τ-step-ahead prediction results for experiments with CISD dataset. Shown: RMSE as mean ± standard deviation over ten runs.

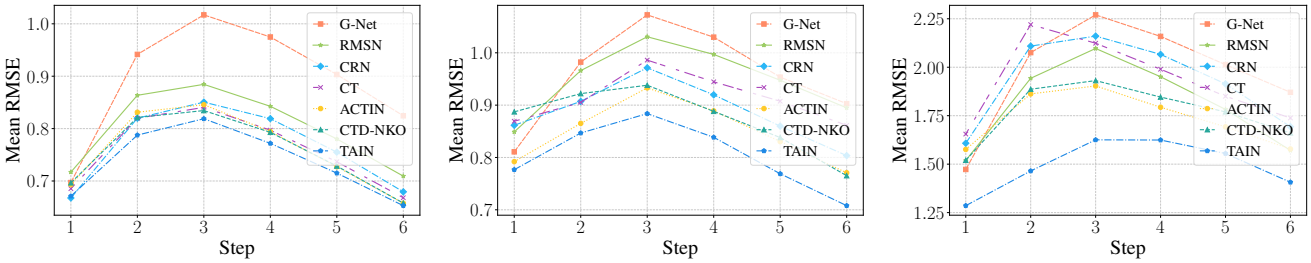
In the original dataset, both radiotherapy and chemotherapy interventions are binary. However, in real-world clinical settings, these treatments often involve varying dosages. To better reflect this complexity, we have adapted the interventions to be continuous rather than binary. This modification allows for a more nuanced representation of treatment intensities, aligning our simulations more closely with clinical realities. For a detailed description of the dataset generation process, please refer to Appendix D.

Results. Figure 3 illustrates the performance comparison between TAIN and the baseline models on the FS-Tumor dataset. The results demonstrate that TAIN exhibits superior or highly competitive performance across different levels of time-varying confounding (γ = 0, 2, and 4) and prediction steps (1 to 6 steps). In most cases, TAIN and the baseline models share a similar performance trend, typically reaching a peak RMSE around the 3rd step and then showing a decreasing trend in long-term predictions. It is noteworthy that this trend does not stem from a reduction in the difficulty of long-term prediction tasks, but rather may be attributed to the reduction of tumor size after the application of treatment interventions, leading to a decrease in RMSE.

It is especially important to note that as confounding bias becomes more severe, treatment allocation grows increasingly imbalanced, complicating the capture of temporal treatment interactions. In such scenarios, the benefits of TAIN are pronounced, demonstrating the superiority of its distinctive IFC approach. For more comprehensive experimental results on the FS-Tumor dataset, please refer to Appendix D.

5.2 Experiments with CISD Data

Data. The Continuous Intervention Synthetic Dataset (CISD) [Wang *et al.*, 2024] is a synthetic time series dataset designed to simulate the effects of continuous interventions. In this dataset, the generation of treatment variables relies on non-



(a) $\gamma = 0$ (b) $\gamma = 2$ (c) $\gamma = 4$

Figure 3: Performance comparison of the TAIN model with alternative models for 1-step to 6-step predictions on FS-Tumor datasets. Results are presented for three levels of time-varying confounding factor: $\gamma = 0$, $\gamma = 2$, and $\gamma = 4$. Each subplot shows the mean RMSE over ten runs for a specific γ value, with the x-axis representing the prediction steps (1 to 6) and the y-axis showing the RMSE.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
ACTIN	1.58 ± 0.47	$\mathbf{1.86 \pm 0.92}$	$\mathbf{1.90 \pm 1.02}$	1.79 ± 0.97	1.69 ± 0.94	1.58 ± 0.89
ACTIN w/o balancing	1.58 ± 0.47	1.88 ± 0.99	1.92 ± 1.06	1.79 ± 0.97	$\mathbf{1.67} \pm 0.91$	$\mathbf{1.54} \pm 0.84$
TAIN	1.29 ± 0.50	$\mathbf{1.46 \pm 0.65}$	$\mathbf{1.63 \pm 0.78}$	$\mathbf{1.62 \pm 0.81}$	$\mathbf{1.55 \pm 0.81}$	$\mathbf{1.41 \pm 0.76}$
TAIN w/o DG	$\mathbf{1.26 \pm 0.47}$	1.53 ± 0.70	1.69 ± 0.84	1.68 ± 0.87	1.61 ± 0.87	1.47 ± 0.82

Table 2: Ablation study results for ACTIN and TAIN on the FS-Tumor dataset with $\gamma = 4$

linear transformations of historical covariate data and the addition of noise, reflecting confounding bias through a Beta distribution. Furthermore, the generation of covariate and outcome variables significantly depends on complex nonlinear transformations that not only capture the dynamic influence of historical covariates and treatment data but also enhance the realism of the dataset. These nonlinear processing techniques make the CISD dataset a useful tool for investigating the intricate dynamic effects of continuous interventions, making it particularly suitable for simulating treatment effects in real-world scenarios. For a detailed description of the dataset generation process, please refer to Appendix D.

Results. Table 1 presents the τ-step prediction results on the CISD dataset. Compared to the FS-Tumor dataset, CISD has a higher dimensionality of covariates, resulting in more complex temporal treatment interactions. As the number of prediction steps increases, this complexity leads to greater challenges in making accurate predictions, which can be observed from the performance of different models. The results demonstrate that TAIN consistently outperforms other baseline models across all prediction steps. As the number of prediction steps increases, TAIN shows a clearer advantage over baseline models, demonstrating its effectiveness in capturing temporal treatment interactions through the IFC approach. In contrast, models like ACTIN treat treatments as general inputs, limiting their ability to model nonlinear complexities.

5.3 Ablation Studies

To further investigate the effectiveness of our proposed theory, we conducted an ablation study on the FS-Tumor dataset with $\gamma = 4$. We compared the performance of the TAIN model with and without the domain generalization strategy to evaluate the impact of the proposed theory on alleviating confounding bias. To contrast with balancing strategies from the perspective of domain adaptation, we also selected ACTIN as a representative for the ablation experiment.

The ablation study results demonstrate the effectiveness of the domain generalization (DG) strategy in TAIN for addressing confounding bias. As shown in Table 2, TAIN with DG consistently outperforms its counterpart without DG for longer time steps ($\tau \geq 2$), with improvements ranging from 0.06 to 0.07 across $\tau = 2$ to $\tau = 6$. In contrast, while ACTIN also shows improvements with its balancing component, the gains are relatively smaller compared to TAIN's DG strategy. This comparison highlights that our proposed domain generalization approach is more effective at handling the distribution shift between training and testing phases, particularly when dealing with confounding bias. The superior performance of TAIN can be attributed to its explicit consideration of the unique challenge in counterfactual estimation: the unavailability of target domain distributions during testing, which is directly addressed through our framework. Additional analyses on the SAM radius ρ and the RBF-based treatment representation are provided in Appendix F.

6 Conclusion

In this paper, we propose TAIN for longitudinal counterfactual estimation under confounding bias and temporal complexity. By encoding treatment information into the IFC kernel, TAIN highlights treatment effects and captures temporal treatment interactions. It also mitigates time-varying confounding bias by bounding counterfactual estimation error from a domain generalization perspective. Theoretical and empirical results demonstrate its effectiveness, with future work extending evaluation to real-world datasets.

Ethical Statement

This study requires no institutional ethics approval. All results are based on simulated and publicly available datasets.

Acknowledgments

This research was supported in part by the National Nature Science Foundation of China (No. 62137002, 62406302, 62576327), in part by the Natural Science Foundation of Anhui province (No. 2408085QF195), in part by the USTC Research Funds of the Double First-Class Initiative (Grant No. YD9110002085), in part by the Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (Grant No. 2025IHM01030), and, in part, by Jinan-NTU Green Technology Research Institute (GreenTRI).

References

[Bai *et al.*, 2018] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[Ban *et al.*, 2025a] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen. Llm-driven causal discovery via harmonized prior. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Ban *et al.*, 2025b] Taiyu Ban, Changxin Rong, Xiangyu Wang, Lyuzhou Chen, Xin Wang, Derui Lyu, Qinrui Zhu, and Huanhuan Chen. Differentiable structure learning with ancestral constraints. In *Forty-second International Conference on Machine Learning*, 2025.

[Bica *et al.*, 2020a] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International conference on machine learning*, pages 884–895. PMLR, 2020.

[Bica *et al.*, 2020b] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

[Chen *et al.*, 2023] Yehu Chen, Annamaria Prati, Jacob Montgomery, and Roman Garnett. A multi-task gaussian process model for inferring time-varying treatment effects in panel data. In *International Conference on Artificial Intelligence and Statistics*, pages 4068–4088. PMLR, 2023.

[Chen *et al.*, 2025] Lyuzhou Chen, Taiyu Ban, Derui Lyu, Yijia Sun, Kangtao Hu, Xiangyu Wang, and Huanhuan Chen. Continuous structure constraint integration for robust causal discovery. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[Chevillon, 2007] Guillaume Chevillon. Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785, 2007.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 103. Association for Computational Linguistics, 2014.

[Foret *et al.*, 2021] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[Frauen *et al.*, 2023] Dennis Frauen, Tobias Hatt, Valentyn Melnychuk, and Stefan Feuerriegel. Estimating average causal effects from patient trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7586–7594, 2023.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[Geng *et al.*, 2017] Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Scientific reports*, 7(1):13542, 2017.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Hariton and Locascio, 2018] Eduardo Hariton and Joseph J Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

[Hassanpour and Greiner, 2019] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hess *et al.*, 2024] Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024.

[Huang and Ning, 2012] Xuelin Huang and Jing Ning. Analysis of multi-stage treatments for recurrent diseases. *Statistics in medicine*, 31(24):2805–2821, 2012.

[Huang *et al.*, 2024] Qiang Huang, Chuizheng Meng, Defu Cao, Biwei Huang, Yi Chang, and Yan Liu. An empirical examination of balancing strategy for counterfactual estimation on time series. In *Forty-first International Conference on Machine Learning*, 2024.

[Johansson *et al.*, 2016] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[Johansson *et al.*, 2022] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *The Journal of Machine Learning Research*, 23(1):7489–7538, 2022.

[Li *et al.*, 2021] Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 282–299. PMLR, 04 Dec 2021.

[Lim *et al.*, 2018] Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

[Matsuura and Harada, 2020] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11749–11756, 2020.

[Melnychuk *et al.*, 2022] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15329. PMLR, 2022.

[Meng *et al.*, 2023] Chuizheng Meng, Yihe Dong, Sercan Ö Arık, Yan Liu, and Tomas Pfister. Costar: Improved temporal counterfactual estimation with self-supervised learning. *arXiv preprint arXiv:2311.00886*, 2023.

[Motiian *et al.*, 2017] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

[Nie *et al.*, 2020] Lizhen Nie, Mao Ye, Dan Nicolae, et al. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *Proceedings of the 9th International Conference on Learning Representations*, 2020.

[Oord *et al.*, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Robins and Hernan, 2008] James Robins and Miguel Hernan. Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599, 2008.

[Robins, 1986] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[Roemhild *et al.*, 2022] Roderich Roemhild, Tobias Bollenbach, and Dan I Andersson. The physiology and genetics of bacterial responses to antibiotic combinations. *Nature Reviews Microbiology*, 20(8):478–490, 2022.

[Rubin, 1978] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[Seedat *et al.*, 2022] Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *International Conference on Machine Learning*, pages 19497–19521. PMLR, 2022.

[Shalit *et al.*, 2017] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[Shi *et al.*, 2021] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[Taieb and Atiya, 2015] Souhaib Ben Taieb and Amir F Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE transactions on neural networks and learning systems*, 27(1):62–76, 2015.

[Wang *et al.*, 2024] Xin Wang, Shengfei Lyu, Lishan Yang, Yibing Zhan, and Huanhuan Chen. A dual-module framework for counterfactual estimation over time. In *Forty-first International Conference on Machine Learning*, 2024.

[Wang *et al.*, 2025] Xin Wang, Shengfei Lyu, Kangyang Luo, Lishan Yang, Huanhuan Chen, and Chunyan Miao. Enhancing counterfactual estimation: a focus on temporal treatments. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6442–6450, 2025.

[Xu *et al.*, 2020] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6502–6509, 2020.

[Zhang *et al.*, 2024] Ruipeng Zhang, Ziqing Fan, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Domain-inspired sharpness-aware minimization under domain shifts. In *The Twelfth International Conference on Learning Representations*, 2024.