Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

GANCQR: ESTIMATING PREDICTION INTERVALS FOR INDIVIDUAL TREATMENT EFFECTS WITH GANS

Jiaxing Wang[1], Hong Wan[2], and Xi Chen[3]

[1]Operations Research Graduate Program, North Carolina State University, Raleigh, NC, USA
[2]Edward P. Fitts Dept. of Industrial and Systems Eng., North Carolina State University, Raleigh, NC, USA
[3] Grado Dept. of Industrial and Systems Eng., Virginia Tech, Blacksburg, VA, USA

ABSTRACT

Evaluating individual treatment effects (ITE) is challenging due to the lack of access to counterfactual outcomes, particularly when working with biased data. Recent efforts have focused on leveraging the generative capabilities of models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) for ITE estimation. However, few approaches effectively address the need for uncertainty quantification in these estimates. In this work, we introduce GANCQR, a GAN-based conformal prediction method that generates prediction intervals for ITE with reliable coverage. Numerical experiments on synthetic and semi-synthetic datasets demonstrate GANCQR's superiority in handling selection bias compared to state-of-the-art methods.

1 INTRODUCTION

Counterfactual inference is a method focusing on estimating the outcomes of a system under different conditions that were not actually experienced. It involves constructing a counterfactual or "what-if" scenario to estimate what the outcome would have been if a different treatment or intervention had been assigned. This method has received attention from a variety of fields, including political phenomena (Grimmer et al. 2017), clinical management and precision medicine (Prosperi et al. 2020; Foster et al. 2011; Ge et al. 2020), policy evaluation (Chernozhukov et al. 2021), and economics (Florens et al. 2008). In particular, counterfactual inference can be used to assess the effect of a treatment in a medical study by comparing the actual outcome with the estimated counterfactual outcome. Estimating treatment effects differs from standard supervised learning problems because counterfactuals are hypothetical, making it impossible to observe individual-level effects directly. Early studies concentrated on the average treatment effect (ATE) to capture population-level difference (Cochran and Rubin 1973; Rubin 2005). However, ATE only provides the central tendency of the treatment effect distribution, making it insufficient to validate a treatment's effectiveness, especially for non-randomized data. As a remedy, the average treatment effect on the treated group (ATT) was introduced to estimate the impact on the treated group in non-randomized data (Heckman and Robb 1985; Heckman, Ichimura, and Todd 1997). Since the early 2010s, machine learning models have significantly influenced healthcare (Xiao et al. 2018). Hospitals and medical providers increasingly embrace machine learning methods for clinical decision-making to maintain or improve an individual's health. This trend has spurred the proposal of individual treatment effects (ITE) to support individualized patient care. Due to challenges in ITE estimation, most studies estimate conditional average treatment effects (CATE), representing ITE expectations conditioned on covariate values. While CATE provides enhanced insight over ATE, it still neglects the inherent variability in the response, often referred to as conditional variance (Lei and Candès 2021). This variability might be crucial for decision-making, especially if the covariates fail to explain most of the variation in ITE.

The major challenge in ITE estimation is the absence of counterfactual data, which represents the unobserved outcomes of the compensate treatments (Hill 2011). Randomized Controlled Trials (RCTs)

2024 Winter Simulation Conference (WSC) | 979-8-3315-3420-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/WSC63780.2024.10838943

offer a solution by randomly assigning participants to treatment or control groups to compensate for missing counterfactuals (Haynes et al. 2012). Despite being considered the golden standard for effective causal inference methods, RCTs suffer from multiple issues, including high cost, relatively small size, ethical issues, and short duration of followups that might miss out on long-term effects of medications (Ghosheh et al. 2024). Moreover, treatment allocation is not randomized in observational datasets like electronic health records (EHRs) encompassing covariates, treatment assignments, and corresponding outcomes. Consequently, this non-random nature may result in biased control group selection. The generative capabilities of GANs offer a promising option to address this issue for treatment effect estimation by generating missing counterfactual outcomes, adjusting the distribution of observed data, or augmenting original sample size (Ghosheh et al. 2024). Compared to VAEs, which learn a balanced representation to estimate ITE, GANs do not have to trade off between containing predictive information and reducing biased information (Yoon et al. 2018).

Point estimates can be risky for decision-making in contexts like drug development, where incorrect actions can lead to significant losses. In such cases, a confidence interval becomes crucial. For instance, the U.S. FDA requires an interval estimation or at least a p-value for drug approval to ensure sufficient evidence and confidence in the drug. Despite the importance of uncertainty quantification, reliable methods for constructing prediction intervals for ITE still need to be developed. Lei and Candès (2021) demonstrated the unsatisfactory or unacceptable coverage of the confidence intervals for CATE and prediction intervals for ITE generated by widely used methods, including Bayesian approaches, even in elementary models with only ten covariates.

The literature on ITE estimation is still in development, with only a limited number of studies having been conducted. In this paper, we focus solely on methods that leverage the generative capabilities of GANs, although other methods exist that take advantage of Variational Autoencoders (VAEs). Yoon et al. (2018) introduced GANITE, the first application of GANs for inferring ITE. GANITE treats counterfactuals as missing labels and imputes the missing data adapted from GANIN proposed in their previous work (Yoon et al. 2018). Additionally, McDermott et al. (2018) proposed a cycle Wasserstein regression GAN model to generate time-series post-treatment outcomes for biomedical applications to predict a patient's response to a treatment. Inspired by GANITE, Ge et al. (2020) extended GANITE to MGANITE to estimate ITE for all types of treatments including binary, categorical, and continuous treatments. Furthermore, Generative Adversarial De-confounding (GAD) is proposed to deal with the continuous treatment by Kuang et al. (2021). GAD generates a "calibration" distribution to eliminate the associations between covariates and treatment variables. In this process, GANs are employed to learn a weight vector that transfers the distribution of observed data to the "calibration" distribution.

The only research on constructing a prediction interval for ITE has been conducted by Lei and Candès (2021). They calculated prediction intervals for ITE by targeting estimating prediction intervals of the potential outcomes individually first. However, separately estimating the potential outcomes can be inefficient and unreliable, especially when treatments are not assigned completely randomly (Curth, Peck, McKinney, Weatherall, and van Der Schaar 2024). Various work has shown that it is more efficient to estimate ATE and CATE directly (Van der Laan and Rose 2011; Künzel et al. 2019; Athey et al. 2019; Nie and Wager 2021). GANs in this case can generate the counterfactual data directly, therefore allowing direct estimation of ITE, analogous to the approaches in estimating ATE and CATE.

In this paper, we construct prediction intervals for ITE directly by leveraging the ability of GANs and conformal prediction. In particular, we address two challenges: (i) constructing prediction intervals for ITEs with reliable coverage without prior knowledge of the distributions of potential outcomes, and (ii) constructing efficient prediction intervals for ITEs when the observed data exhibit selection bias. The rest of the paper is organized as follows. Section 2 sets up the theoretical framework for ITE estimation. Section 3 presents GANCQR in detail. Section 4 provides numerical experiments based on semi-synthetic and synthetic datasets. Section 5 concludes the paper.

2 PROBLEM FORMULATION

In this paper, we focus on the standard potential outcome setting (Rubin 2005) with binary treatment to estimate the treatment effects. Let $X \in \mathscr{X} \subset \mathbb{R}^s$ denote the observed $s-$dimensional feature vector, $T \in \{0,1\}$ denote the binary treatment indicator, and $(Y(0),Y(1)) \in \mathscr{Y}^2$ represent the pair of potential outcomes where $\mathscr{Y}$ is the set of possible outcomes. Suppose the joint distribution of $(X,T,Y(0),Y(1))$ is μ; the marginal distribution of X is denoted by μ_X, and the conditional marginal distribution of $\mathbf{Y} \triangleq (Y(0),Y(1))$ conditioning on X is denoted by $\mu_{\mathbf{Y}}(X)$. Assume that there are N i.i.d. observed samples $(\mathbf{x}_i,T_i,Y_i^{obs})$ for $i = 1,2,\ldots,N$, where only one treatment t is assigned for each sample $\mathbf{x}$, resulting in only one potential outcome to be observed. The Y_i^{obs} and ITE are defined under three commonly stipulated assumptions: the overlap assumption, the unconfoundedness assumption, and the stable unit treatment value assumption (SUTVA).

Assumption 1 (**Overlap**) For $\forall \mathbf{x} \in \mathscr{X}$, we have $0 < P(T = 1 \mid X = \mathbf{x}) < 1$.

The overlap assumption ensures the probability of receiving treatment is positive for every point in the covariate space.

Assumption 2 (**Unconfoundedness**) Conditional on X, the potential outcomes $(Y(0),Y(1))$ are independent of T, i.e. $(Y(0),Y(1)) \perp T \mid X$.

Unconfoundedness, also known as strong ignorability, allows covariates in X directly affect the values of $Y(1)$ and $Y(0)$, but the treatment status is unrelated to these values. It rules out any sources of unmeasured confounders, ensuring that individuals who receive the treatment are not fundamentally different from those who do not. This assumption enables us to study what $Y(0)$ would have been for those individuals with $T = 1$ by looking at the effect of those with $T = 0$.

Assumption 3 (**SUTVA**) SUTVA consists of two elements: no interference and no hidden variations of a treatment (Imbens and Rubin 2015). No interference means that one observation's potential outcome is not affected by the treatment assignments of other observations. No hidden variations refer to the fact that, for each observation, there are no different forms or versions of each treatment level that would lead to different potential outcomes.

Then the observed outcome Y_i^{obs} represents the factual outcome, denoted as y_f, which refers to the component of the potential outcomes that corresponds to the assigned treatment $Y_i^{obs} = T_i \cdot Y_i(1) + (1-T_i) \cdot Y_i(0)$. The unobserved potential outcomes corresponding to the opposite treatment assignment are defined as counterfactual outcomes or counterfactual in short denoted as y_{cf}. The joint distribution of observed sample (X,T,y_f) is denoted by μ_f. The individual treatment effect τ is defined as

$$\tau = Y(1) - Y(0) \ . \tag{1}$$

We want to conduct a prediction interval for τ, denoted by $\hat{C}_{\text{ITE}}(X)$, which satisfies the coverage criterion $P_{(X,\tau)}(\tau \in \hat{C}_{\text{ITE}}(\mathbf{x})) \geq 1-\alpha$ for a pre-specified level $\alpha \in (0,1)$.

3 GANCQR

This section begins by introducing GANITE and conformalized quantile regression (CQR). We then delve into GANCQR which combines GRANITE and CQR to construct a covariate-dependent prediction interval for ITE.

3.1 GANITE Model

GANITE is the first algorithm that uses GANs for ITE inference (Yoon et al. 2018). It utilizes a pair of conditional GANs (Mirza and Osindero 2014): one for the counterfactual imputation (via the counterfactual block with a counterfactual generator $\mathbf{G}$ and a discriminator $\mathbf{D_G}$), and another for ITE estimation (via the ITE block with an ITE generator $\mathbf{I}$ and a discriminator $\mathbf{D_I}$). GANs take a random vector $\mathbf{z}$ as input and

use deep learning methods to generate new samples in a given domain. Conditional GANs extend GANs by conditioning both the generator and the discriminator on additional information.

The counterfactual block focuses on generating samples from the distribution of potential outcomes $\mu_{\mathbf{Y}}(\mathbf{x})$ for any given $\mathbf{x} \in \mathscr{X}$. However, the absence of counterfactuals makes it impossible to learn $\mu_{\mathbf{Y}}(\mathbf{x})$ directly. Therefore, the counterfactual generator $\mathbf{G}$ estimates $\tilde{\mathbf{y}} = (\tilde{y}_f, \tilde{y}_{cf})$ from the conditional distribution of $(y_f, y_{cf}) \stackrel{\triangle}{=} \mathbf{Y}_{cf}$ given $X = \mathbf{x}, T = t$ and $Y^{obs} = y_f$. This conditional distribution is denoted by $\mu_{\mathbf{Y}_{cf}}(\mathbf{x}, t, y_f)$. The counterfactual generator $\mathbf{G}(\mathbf{x}, t, y_f)$ is defined as $\mathbf{G}(\mathbf{x}, t, y_f) = g(\mathbf{x}, t, y_f, \mathbf{z}_{\mathbf{G}})$, where g is a function $g : \mathscr{X} \times \{0,1\} \times \mathscr{Y} \times [-1,1] \to \mathscr{Y}^2$ and $\mathbf{z}_{\mathbf{G}} \sim \text{Uniform}(-1,1)$. One needs to find a suitable g such that $\mathbf{G}(\mathbf{x}, t, y_f)$ approximates the target distribution $\mu_{\mathbf{Y}}(\mathbf{x}, t, y_f)$. Let $\bar{\mathbf{y}} = (y_f, \tilde{y}_{cf})$ be the vector obtained by replacing $\tilde{y}_\eta$ with y_f, which means the ηth component of $\tilde{\mathbf{y}}$ is y_f. The corresponding discriminator $\mathbf{D}_{\mathbf{G}}$ intends to identify which component of $\tilde{\mathbf{y}}$ represents the factual outcome. Let $\mathbf{D}_{\mathbf{G}}(\mathbf{x}, \tilde{\mathbf{y}}) \in [0,1]$ representing the probability of correctly identifying the factual outcome. In the counterfactual block, we train $\mathbf{D}_{\mathbf{G}}$ to maximize the probability of correctly identifying η and train $\mathbf{G}$ to minimize the probability of $\mathbf{D}_{\mathbf{G}}$ correctly identifying η by solving the following minimax problem:

$$\min_{\mathbf{G}} \max_{\mathbf{D}_{\mathbf{G}}} \mathbb{E}_{(\mathbf{x},t,y_f)\sim\mu_f} \left(\mathbb{E}_{\mathbf{z}_{\mathbf{G}}\sim\mathscr{U}(-1,1)} \left[t \log \mathbf{D}_{\mathbf{G}}(\mathbf{x}, \tilde{\mathbf{y}}) + (1-t) \log(1 - \mathbf{D}_{\mathbf{G}}(\mathbf{x}, \tilde{\mathbf{y}})) \right] \right).$$

GANITE adopts minibatches to avoid mode collapse by discriminating between k_G minibatches of samples, rather than between individual samples. To enforce the estimated factual outcome $\tilde{y}_f$ identified by $\mathbf{D}_{\mathbf{G}}$ equals the factual outcome y_f, a supervised loss, denoted by $L_S^{\mathbf{G}}$, is added as a constraint. For a given sample, $L_S^{\mathbf{G}}(y_f, \tilde{y}_\eta) = (y_f - \tilde{y}_\eta)^2 \in \mathbb{R}$, where $y_f \in \mathbb{R}$ and $\tilde{y}_\eta \in \mathbb{R}$ is the ηth component in $\tilde{\mathbf{y}}$ that is identified by $\mathbf{D}_{\mathbf{G}}$ as the factual outcome. The empirical objective function of the minimax problem in the counterfactual block is defined by $V_{CF}(\mathbf{x}, t, \tilde{\mathbf{y}}) = t \log(\mathbf{D}_{\mathbf{G}}(\mathbf{x}, \tilde{\mathbf{y}})) + (1-t) \log(1 - \mathbf{D}_{\mathbf{G}}(\mathbf{x}, \tilde{\mathbf{y}})) \in \mathbb{R}$, where $\mathbf{x} \in \mathbb{R}^s$, $t \in \mathbb{R}$, and $\tilde{\mathbf{y}} \in \mathbb{R}^2$. With the above two objective functions and given the number of minibatches k_G, we iteratively optimize $\mathbf{G}$ and $\mathbf{D}_{\mathbf{G}}$ as follows:

$$\begin{aligned} &\min_{\mathbf{D}_{\mathbf{G}}} - \sum_{n=1}^{k_G} V_{CF}(\mathbf{x}, t, \tilde{\mathbf{y}}) \\ &\min_{\mathbf{G}} \sum_{n=1}^{k_G} \left[V_{CF}(\mathbf{x}, t, \tilde{\mathbf{y}}) + \alpha_{\mathbf{G}} L_S^G(y_f, \tilde{y}_\eta) \right], \end{aligned} \tag{2}$$

where $\alpha_{\mathbf{G}} \geq 0$ is a pre-determined hyper-parameter. After training the counterfactual generator, the estimated factual outcome $\tilde{y}_f$ is replaced with the original factual outcome to form a complete dataset $\tilde{D}_i = \{\mathbf{x}_i, t_i, \bar{\mathbf{y}}_i\}$. Then we pass on the complete dataset $\tilde{\mathbf{D}} = \left\{\tilde{D}_i\right\}$ to the ITE block.

In the ITE block, the ITE generator $\mathbf{I}$ uses the covariates $\mathbf{x}$ to estimate the potential outcomes denoted by $\hat{\mathbf{y}}$. Here the function h should best approximate the marginal distribution of potential outcomes $\mu_{\mathbf{Y}}(\mathbf{x})$, where h is a function $h : \mathscr{X} \times [-1,1] \to \mathscr{Y}^2$ and $\mathbf{z}_{\mathbf{I}} \sim \text{Uniform}(-1,1)$, $\mathbf{I}(\mathbf{x})$ is defined as $\mathbf{I}(\mathbf{x}) = h(\mathbf{x}, \mathbf{z}_{\mathbf{I}})$.. Then we can compute the ITE using the estimated $\hat{\mathbf{y}}$ according to (1). The ITE discriminator $\mathbf{D}_{\mathbf{I}}$ adapts the conditional GANs discriminator which distinguishes between the synthetic data $\hat{\mathbf{y}}$ and the data $\mathbf{y}^*$ from the complete data $\tilde{D}$ obtained by the counterfactual block. The minimax problem that formulates this procedure is given by

$$\min_{\mathbf{I}} \max_{\mathbf{D}_{\mathbf{I}}} \mathbb{E}_{\mathbf{x}\sim\mu_{\mathbf{X}}} \left(\mathbb{E}_{y^*\sim\mu_{\mathbf{Y}}(\mathbf{x})} \left[\log \mathbf{D}_{\mathbf{I}}(\mathbf{x}, \mathbf{y}^*) \right] + \mathbb{E}_{y^*\sim\mathbf{I}(\mathbf{x})} \left[\log(1 - \mathbf{D}_{\mathbf{I}}(\mathbf{x}, \mathbf{y}^*)) \right] \right),$$

where $\mathbf{D}_{\mathbf{I}}(\mathbf{x}, \mathbf{y}^*) \in [0,1]$ is the probability of $\mathbf{y}^*$ coming from $\tilde{D}$. To obtain a smaller mean square error for the estimated ITE, a supervised loss $L_S^{\mathbf{I}}(\mathbf{y}^*, \hat{\mathbf{y}}) = ((\mathbf{y}_1^* - \mathbf{y}_0^*) - (\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0))^2 \in \mathbb{R}$ is added for the ITE block, where $\mathbf{y}_1^* \in \mathbb{R}$ is the potential outcome with $t = 1$ obtained from the complete data $\tilde{D}$ ($\mathbf{y}_0^*$ is the potential outcome with $t = 0$ respectively) and $\hat{\mathbf{y}}_1 \in \mathbb{R}$ is the potential outcome with $t = 1$ generated by $\mathbf{I}$ ($\hat{\mathbf{y}}_0$ is the

potential outcome with $t = 0$ respectively). The empirical objective function of the minimax problem in the ITE block is defined by $V_{ITE}(\mathbf{x}, \mathbf{y}^*, \hat{\mathbf{y}}) = \log(\mathbf{D_I}(\mathbf{x}, \mathbf{y}^*)) + \log(1 - \mathbf{D_I}(\mathbf{x}, \hat{\mathbf{y}})) \in \mathbb{R}$, where $\mathbf{x} \in \mathbb{R}^s$, $\mathbf{y}^* \in \mathbb{R}^2$, and $\hat{\mathbf{y}} \in \mathbb{R}^2$. With the above two objective functions, given the number of minibatch k_I, $\mathbf{I}$ and $\mathbf{D_I}$ are iteratively optimized as follows:

$$\begin{aligned} &\min_{\mathbf{D_I}} - \sum_{n=1}^{k_I} V_{ITE}(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}}) \\ &\min_{\mathbf{I}} \sum_{n=1}^{k_I} \left[V_{ITE}(\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}}) + \beta_{\mathbf{I}} L_S^I(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) \right], \end{aligned} \tag{3}$$

where $\beta_{\mathbf{I}} \geq 0$ is a pre-determined hyper-parameter. The stochastic gradient descent method is used to solve the optimization problems (2) and (3).

3.2 Conformalized Quantile Regression (CQR)

Conformal prediction (CP) (Vovk et al. 2005; Shafer and Vovk 2008) is a framework for distribution-free uncertainty quantification for regression and classification problems. CP takes an arbitrary prediction model as the input and produces a prediction interval with guaranteed coverage. Given an arbitrary method for making a prediction $\hat{y}$ of a label y and a significance level (or target miscoverage level) $\alpha \in (0,1)$, CP produces a prediction set $\hat{C}$ containing both $\hat{y}$ and y with probability at least $1 - \alpha$, where $\hat{y}$ denotes the point prediction, and $\hat{C}$ is referred to as the region prediction. The only assumption stipulated by CP is exchangeability. Given a measurable space $\mathbf{Z}$ and variables $\{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^N \in \mathbf{Z}^N$, the definition of exchangeability is given as follows.

Definition 1 If for any permutation π of integers $1, 2, ..., N$, the variables $w_1, ..., w_N$ have the same joint probability distribution as $z_1, ..., z_N$ where $w_i = z_{\pi(i)}$, then the variables $z_1, ..., z_N$ are exchangeable.

We use the term bag, denoted by $B = \wr z_1, ..., z_N \wr \in \mathbf{Z}^N$, to represent the collection of elements obtained from a list $z_1, ..., z_N$ by removing information about the ordering. Starting from a point prediction method γ, we can define a real-valued function $A(B, z) : \mathbf{Z}^N \times \mathbf{Z} \to \mathbb{R}$, referred to as the nonconformity measure, which quantifies how unusual an example $z \in \mathbb{R}$ looks compared to the examples in the bag B. Given a metric that measures the distance $d(z, z')$ between two elements, A is defined by $A(B, z) = d(\gamma(B), z)$. Given a nonconformity measure A, the nonconformity score is defined as $E = A(B, z)$. The prediction interval given by CP is the collection of all z such that $\{z \in \mathbf{Z} : p_z > \alpha\}$ where $p_z = |\{i = 1, 2, \ldots, N+1 : E_i \geq E_{N+1}\}| / (N+1)$, and $E_i = A(\wr z_1, ..., z_N, z \wr \setminus \wr z_i \wr, z_i)$ for $i = 1, 2, \ldots, N$, $E_{N+1} = A(\wr z_1, \ldots, z_N \wr, z)$.

Split conformal prediction splits the original data into two disjoint subsets: a proper training set $\{z_i = (\mathbf{x}_i, y_i) : i \in \mathscr{I}_1\}$ and a calibration set $\{z_i = (\mathbf{x}_i, y_i) : i \in \mathscr{I}_2\}$. This division ensures that the nonconformity scores are constructed symmetrically, thereby guaranteeing the exchangeability of these scores. Split CQR is a combination of split conformal prediction and quantile regression, which provides finite sample coverage guarantees. (Romano et al. 2019). To compute the nonconformity score, we split the N given data into two subsets: one is the training set indexed by $\mathscr{I}_1$, which is used to obtain two conditional quantile estimators $\hat{q}_{\alpha_{lo}}$, and $\hat{q}_{\alpha_{hi}}$ and the other is the calibration set indexed by $\mathscr{I}_2$, which is used to evaluate the conformity scores E as follows

$$E_i = \max \{\hat{q}_{\alpha_{lo}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{hi}}(X_i)\} \qquad \text{for } i \in \mathscr{I}_2 \ . \tag{4}$$

Then the prediction interval for a given new covariate X_{N+1} is given by

$$C(X_{N+1}) = [\hat{q}_{\alpha_{lo}}(X_{N+1}) - Q_{1-\alpha}(E, \mathscr{I}_2), \hat{q}_{\alpha_{hi}}(X_{N+1}) + Q_{1-\alpha}(E, \mathscr{I}_2)], \tag{5}$$

where $Q_{1-\alpha}(E, \mathscr{I}_2)$ is the $(1-\alpha)(1+|\mathscr{I}_2|^{-1})$ empirical quantile of $\{E_i : i \in \mathscr{I}_2\}$. The prediction interval $C(X_{N+1})$ obtained by CQR satisfies the marginal, distribution-free coverage guarantee.

Theorem 4 If $(X_i, y_i), i = 1, 2, \ldots, N+1$ are exchangeable, then the prediction interval $C(X_{N+1})$ constructed by the split CQR algorithm satisfies $\mathbb{P}\{Y_{N+1} \in C(X_{N+1})\} \geq 1 - \alpha$.

The proof of Theorem 4 can be found in Romano et al. (2019). Notice that Theorem 4 holds uniformly over all conditional distributions $P_{Y|X}$ and all algorithms used to fit conditional quantiles, making CQR a reliable method for counterfactual inference conditioning on covariates.

3.3 GANCQR

We now describe our method GANCQR to construct a prediction interval for ITE τ, as illustrated in Figure 1. Given a dataset $\left\{(\mathbf{x}_i, t_i, y_{f_i})\right\}_{i=1}^{N}$, we first estimate the ITE τ_i conditioning on each $\mathbf{x}_i$ by GANITE. Unlike Lei and Candès (2021) who constructed the prediction intervals for potential outcomes $Y(0)$ and $Y(1)$ separately, GANCQR estimates $\tau = Y(1) - Y(0)$ directly. This helps eliminate the influence of substantial prognostic information, as much of the complexity is mitigated when we directly consider τ. This is especially true given that the potential outcomes $Y(0)$ and $Y(1)$ are complex functions of covariates X, while the treatment effect τ is not. Examples of prognostic information include risk factors that influence outcomes similarly, regardless of treatment status, as well as relatively limited predictive information (Curth et al. 2024). Here we choose GANITE due to its ability to deal with the discrete treatment assignment with the added advantage of being applicable not only to binary treatment but also to multiple treatments, rendering a flexible extension of our algorithm to estimate multiple treatment effects. The generated ITE is denoted by $\hat{\tau}$.

To further conduct prediction interval for ITE $\tau(X)$ conditioning on factors X, we apply CQR to the data $\{(\mathbf{x}_i, \hat{\tau}_i)\}$ obtained by GANITE. We claim that $\{(\mathbf{x}_i, \hat{\tau}_i)\}$ are exchangeable according to de Finetti's representation theorem. De Finetti's representation theorem states that every sequence of conditionally i.i.d. random variables is exchangeable, which was first proposed by de Finetti and extended by Diaconis and Freedman for finite sequences (Diaconis and Freedman 1980). With the definition of ITE generator $\mathbf{I}$, we start from i.i.d. sequence of noise random variable $\left\{\mathbf{z_I}^{(m)}\right\}$ and define for every m $\tau_m = h(\mathbf{x}, \mathbf{Z_I}^{(m)})$ where $\mathbf{x}$ is a random variable independent of $\mathbf{z_I}^{(m)}$ and h is a measurable function, such as a neural network in GANs. Hence, the resulting random sequence $\{\hat{\mathbf{y}}_m, \mathbf{x}\}$ is exchangeable. Therefore, the sequence $\{(\mathbf{x}_i, \hat{\tau}_i)\}$ is also exchangeable since it is a measurable function of an exchangeable sequence.

The use of CQR provides the coverage guarantee conditioning on covariates. Compared with traditional split conformal prediction whose nonconformity score is computed based only on Y, CQR computes the nonconformity score based on the conditional quantiles of $Y \mid X = \mathbf{x}$. Under Assumptions 2 and 3, the joint distribution of (X, Y^{obs}) of the observed treated samples is given by $P_{X|T=1} \times P_{Y(1)|X}$ and the joint distribution of observed untreated samples is given by $P_{X|T=0} \times P_{Y(0)|X}$. Hence the joint distribution of (X, τ) is $P_X \times P_{\tau|X}$ since $\tau = Y(1) - Y(0)$.

With the exchangeable samples $\{(\mathbf{x}_i, \hat{\tau}_i)\}_{i=1}^{N}$ obtained by applying GANITE, we first split this dataset into two disjoint subsets: a training set $\mathscr{D}_1$ and a calibration set $\mathscr{D}_2$. We use the samples in $\mathscr{D}_1$ to train the conditional quantile regression estimators $\hat{q}_{\alpha_{lo}}$ and $\hat{q}_{\alpha_{hi}}$. Then for each sample in $\mathscr{D}_2$, we compute the nonconformity score according to (4). Given a pre-specified significant level $\alpha \in (0, 1)$, we can obtain the prediction interval $\hat{C}(\mathbf{x}_{N+1})$ for a new patient's ITE by (5).

4 NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of GANCQR. We compare our approach with three competing methods, all of which provide prediction intervals of ITEs proposed in Lei and Candès (2021) on two distinct datasets with different types of selection bias. The first is a semi-synthetic benchmark dataset, the Infant Health Development Program (IHDP) dataset, where the potential outcomes are synthesized and therefore fully known. The second dataset is a synthetic generated according to a modified process in Section 3.6 in Lei and Candès (2021).

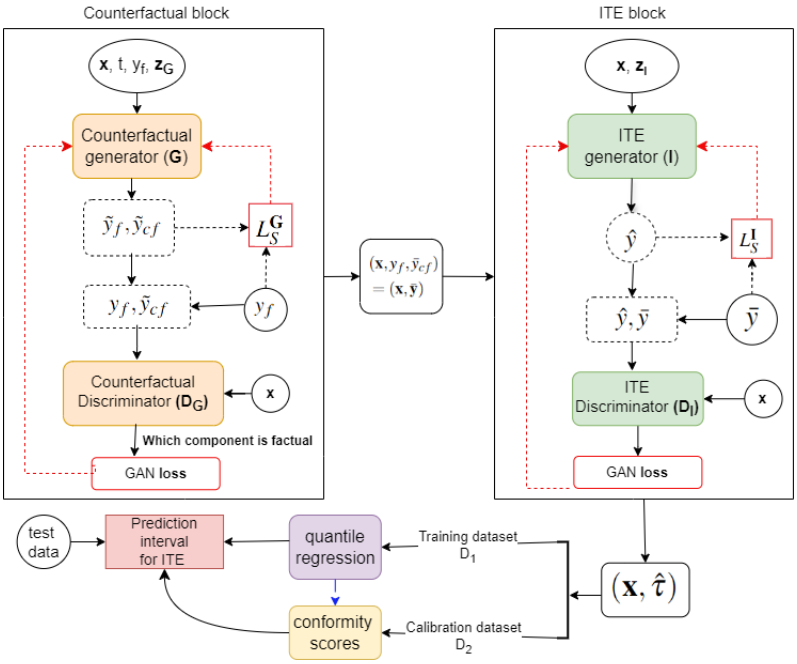


Figure 1: A block diagram of GANCQR.

4.1 Datasets

The IHDP dataset, introduced by Hill (2011), comprises data collected from the Infant Health and Development Program (IHDP), which is designed to predict the effect of receiving specialized childcare on the cognitive test scores of infants. This dataset is commonly used for evaluating treatment effect estimation methods (Louizos et al. 2017; Shalit et al. 2017; Yoon et al. 2018). IHDP consists of 747 units, with 608 control units and 139 treatment units, having 6 continuous and 19 binary covariates. The treatment assignment is "de-randomized" by excluding children with non-white mothers from the treated group. For each sample a treated and a control outcome are simulated following the approach implemented by Shalit et al. (2017), adopting the setting of the response surface type B in Hill (2011). This simulation generates $Y(0)$ according to distribution $\mathcal{N}(\exp((X+W)\beta_B),1)$ and $Y(1)$ according to $\mathcal{N}(X\beta_B-\omega_B^b,1)$, where W is an offset matrix of the same dimension as X with every entry equal to 0.5, and β_B is a $25-$dimensional vector with every entry randomly sampled from $(0,0.1,0.2,0.3,0.4)$ with probabilities $(0.6,0.1,0.1,0.1,0.1)$. For the bth simulation, ω_B^b was chosen so that CATC equals 4 in the incomplete setting where we estimate the effect of the treatment on the controls. Then factual outcomes are then recorded as $Y(1)$ for $t=1$ and $Y(0)$ for $t=0$.

The synthetic dataset satisfies Assumption 2 by generating the covariates $X=[X_1,...,X_d]\sim\mathcal{U}([0,1]^d)$ where d is the dimension of the covariate vector X. The potential outcomes with heteroscedastic errors are generated as $Y(t)\mid X\sim\mathcal{N}(\mu_t(X),\sigma^2(X))$ for $t=0,1$ where $\sigma^2(X))=-\log(X_1)$, $\mu_1(X)=f(X_1)f(X_2)$ and $\mu_0(X)=\gamma f(X_1)f(X_2)$, where $f(x)=1/(1+\exp(-12(x-0.5)))$ and $\gamma=0.5$. The treatment is generated depending on X to create a selection bias according to a propensity score $e(\mathbf{x})=\mathbb{P}(T=1\mid X=\mathbf{x})$ where $\mathbf{x}=(x_1,\cdots,x_s)$. To force that $e(\mathbf{x})\in[0.25,0.5]$, the propensity score is set as $e(\mathbf{x})=(1+\beta_{2,4}(x_1)/4$, where $\beta_{2,4}$ is the cumulative distribution function of the beta distribution with shape parameters $(2,4)$. Then the treatment T is generated as $T\mid X=\mathbf{x}\sim\text{Bern}(e(\mathbf{x}))$. In this example, we consider a low-dimensional case with $d=10$ and 1000 samples.

We follow previous studies (Louizos et al. 2017; Shalit et al. 2017; Yoon et al. 2018) to evaluate our model under two different estimation settings. The first scenario is referred to as the in-sample setting, where we estimate the ITE for all units in a sample with observed factual outcomes for one treatment. This

scenario mimics real-world situations where a cohort is selected once and remains unchanged. Estimating ITE in this context is challenging, as we never observe the ITE for any individual unit. The second scenario is the out-of-sample setting, where we estimate ITE for units without observed outcomes. This corresponds to the challenge of selecting the best treatment for a new patient. Within-sample error is calculated for both the training and validation sets, while out-of-sample error is computed for the test set.

4.2 Benchmark Methods

We compare GANCQR with the state-of-art methods based on weighted split-CQR (WCQR) (Lei and Candès 2021). WCQR takes covariate shift into account and applies weighted conformal prediction since they compute the prediction intervals for $Y(0)$ and $Y(1)$. Take $Y(1)$ for example, the corresponding guaranteed coverage should be $P_{(X,Y)\sim P_{X|T=1}\times P_{Y|X}}(\tau \in \hat{C}_{\text{ITE}}(\mathbf{x})) \geq 1-\alpha$ where the actual distribution of X is $P_{X|T=1} = Q_X$ instead of P_X.

In contrast to our basic idea that first imputes missing counterfactuals to calculate the ITE and then constructs the prediction interval for ITE, Lei and Candès (2021) first derived prediction intervals $\hat{C}_1(\mathbf{x})$ (resp. $\hat{C}_0(\mathbf{x})$) for potential outcomes $Y(1)$ (resp. $Y(0)$) using the samples with $t=1$ (resp. $t=0$). Assume $\hat{C}_0(\mathbf{x}) = \left[\hat{Y}_0^L(\mathbf{x}), \hat{Y}_0^R(\mathbf{x})\right]$ and $\hat{C}_1(\mathbf{x}) = \left[\hat{Y}_1^L(\mathbf{x}), \hat{Y}_1^R(\mathbf{x})\right]$, Lei and Candès (2021) considered three different variants to compute the prediction interval of ITE:The **Naive** method obtains the lower bound by $\hat{Y}_1^L(\mathbf{x}) - \hat{Y}_0^R(\mathbf{x})$ and the upper bound by $\hat{Y}_1^R(\mathbf{x}) - \hat{Y}_0^L(\mathbf{x})$. The **Inexact** Nested method first computes the prediction intervals for ITE τ denoted by $\hat{C}_\tau(\mathbf{x}, t, y_f)$ according to $\hat{C}_\tau(\mathbf{x}, t, y_f) = t(y_f - \hat{C}_0(\mathbf{x})) + (1-t)(\hat{C}_1(\mathbf{x}) - y_f)$. Then compute the desired conditional quantiles of the lower bounds $\hat{C}^L$ and upper bounds $\hat{C}^R$ of $\hat{C}_\tau(\mathbf{x}, t, y_f)$. The **Exact** Nested method follows the same procedure as the Inexact Nested method except for conducting CQR with the induced data set $(\mathbf{x}_i, C_i)$ where $C_i = \left[\hat{C}^L, \hat{C}^R\right]$ instead of fitting their quantiles.

4.3 Performance Evaluation

There are two main indicators to evaluate the performance of interval prediction methods, adaptivity and validity (Angelopoulos and Bates 2023). For validity, we examine whether the coverage meets the coverage requirement. One commonly used metric is the observed prediction interval coverage probability (PICP): $\text{PICP} = N_{test}^{-1} \sum_{i=1}^{N_{test}} \mathbb{1}(\tau_i \in [L_i(X_i), U_i(X_i)])$, where N_{test} is the size of test dataset and τ_i is the ith ITE value. Given a significance level $\alpha \in (0,1)$ and the number of experiments R, the mean of PICP: $\overline{\text{PICP}} = R^{-1} \sum_{k=1}^{R} \text{PICP}_k$ should be approximately $1-\alpha$. However, the observed coverage has fluctuations caused by the finiteness of N, N_{test}, and R. A smaller fluctuation is preferred.

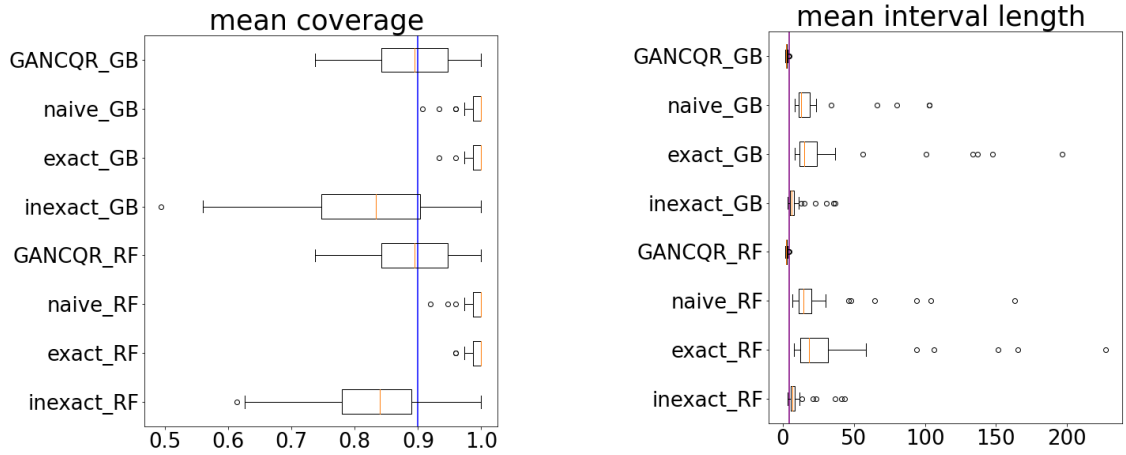
For adaptivity, we check the length of the prediction intervals defined as $|C_i| = U_i(X_i) - L_i(X_i)$ where $C_i = [L_i(X_i), U_i(X_i)]$ is an index of informativeness, and $L_i(X_i)$ and $U_i(X_i)$ are respectively the lower and upper bounds of the prediction interval at X_i. When PICP is acceptable, the smaller the length the better its performance (Fontana et al. 2023). The average prediction interval length is then defined as $\overline{|C|} = N_{test}^{-1} \sum_{i=1}^{N_{test}} |C_i|$. In general, a large $\overline{|C|}$ suggests that the conformal prediction method lacks precision, indicating possible problems in the score or the underlying model.

4.4 Experiment Results

For a fair comparison of all methods, we use 75% samples for training and 25% samples for testing under the out-of-sample setting for all methods. For each testing sample, we produce a $1-\alpha = 0.9$ prediction interval for ITE using the Python library PUNCC (Mendil et al. 2023). We use two different methods to compute the quantile regression: random forest (RF) and gradient boosting (GB). To robustly assess the performance, we conduct 50 independent trials and report PICP and $\overline{|C|}$.

First, we evaluate the performance of GANCQR when tackling the attrition selection bias due to removing non-random subsets of the treated group. Figure 2 presents PICP and $\overline{|C|}$ obtained under the out-of-sample setting and Figure 3 presents those obtained under the in-sample setting. GANCQR exhibits slight

undercoverage since the median of the PICPs obtained by GANCQR is about 0.9. We then calculate $\overline{\text{PICP}}$ for GANCQR using RF and GB as the regression methods and obtain $\overline{\text{PICP}}^{out}_{RF} = 0.8956$, $\overline{\text{PICP}}^{out}_{GB} = 0.8964$, $\overline{\text{PICP}}^{in}_{RF} = 0.8998$, and $\overline{\text{PICP}}^{in}_{GB} = 0.8989$.The resulting $\overline{\text{PICP}}$ values are close to $1-\alpha = 0.9$, with very small fluctuations. Therefore, we find that GANCQR's performance in terms of $\overline{\text{PICP}}$ is acceptable, although some prediction intervals exhibit unsatisfactory coverage. Notice that although the naive and exact methods have satisfactory coverage reflected by the high PICPs, their prediction intervals are too wide to be informative for ITE inference. To evaluate the sharpness of these intervals, we compare the resulting $\overline{|C|}$ with the oracle length. Since the potential outcomes are all normally distributed, the ITE $\tau = Y(1) - Y(0)$ also follows a normal distribution with variance $\sigma_\tau = \sigma^2_{Y(1)} + \sigma^2_{Y(0)} - 2\rho\sigma_{Y(1)}\sigma_{Y(0)}$. Hence, the oracle length is $2 \times 1.645\sigma_\tau \approx 4.237$ since $\sigma_{Y(1)} = \sigma_{Y(0)} = 1$ and the estimated correlation ρ equals 0.3415. As can be seen from Figure 2b and 3b, the $\overline{|C|}$ values corresponding to the naive and exact methods are significantly higher than the oracle length. In contrast, GANCQR delivers the smallest $\overline{|C|}$ values which are close to the oracle length. While the inexact method results in shorter interval lengths, its coverage performance is significantly compromised as a result. Moreover, employing gradient boosting as the regression method for quantile regression outperforms the random forest method, as the former yields smaller $\overline{|C|}$ values while maintaining comparable coverage performance.

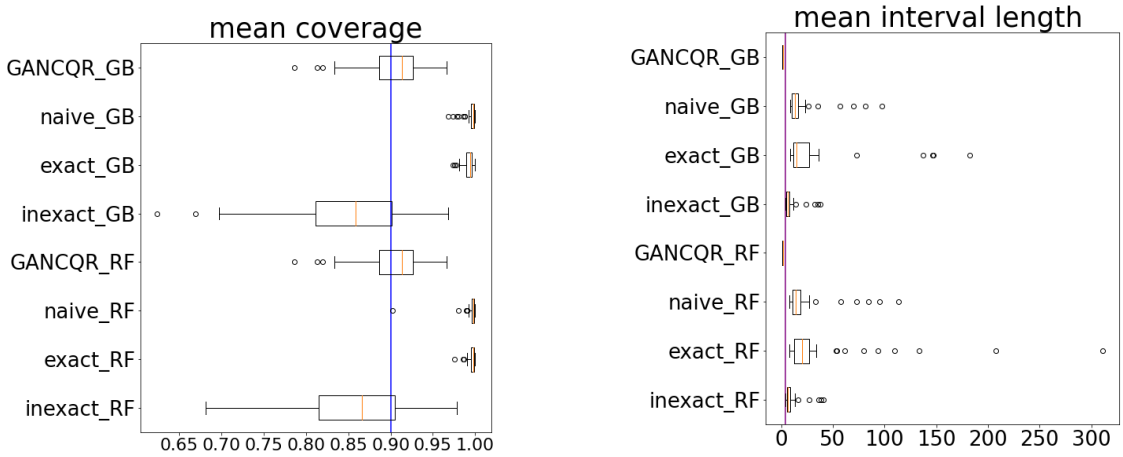


(a) PICP (b) $\overline{|C|}$

Figure 2: IHDP: the results obtained under the out-of-sample setting.

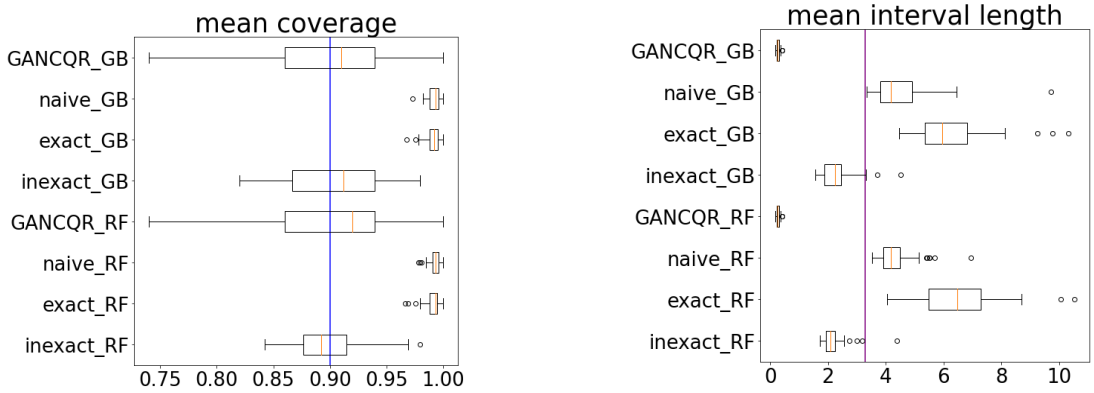
We use the synthetic dataset to examine the performance when tackling the self-selection bias. This bias is induced when the distribution of treatment depends on the features of individuals. For each run, we generate 1000 samples. Figures 4 and 5 present the performance for the synthetic dataset. GANCQR delivers prediction intervals with the smallest $\overline{|C|}$ values, and most of the PICPs obtained by GANCQR are greater than 0.9 since the median is higher than 0.9. We calculate the $\overline{\text{PICP}}$ and obtain $\overline{\text{PICP}}^{out}_{RF} = 0.8995$, $\overline{\text{PICP}}^{out}_{GB} = 0.89947$, $\overline{\text{PICP}}^{in}_{RF} = 0.9029$, and $\overline{\text{PICP}}^{in}_{GB} = 0.90293$. The resulting $\overline{\text{PICP}}$ is almost equal to $1-\alpha = 0.9$. Therefore, the GANCQR's performance in terms of $\overline{\text{PICP}}$ is acceptable. The exact and naive methods both meet the coverage requirement, but the prediction interval lengths are unacceptably large. In this case, the oracle length is $2 \times 1.645\sigma_\tau = 3.29$ as given by Lei and Candès (2021). We see that the $\overline{|C|}$ values corresponding to the inexact and naive methods are too large as they are almost twice the oracle length.

In brief, the results from both in-sample and out-of-sample settings across the two datasets highlight GANCQR's robustness. Its consistent performance suggests potential applicability in crucial real-world scenarios, including selecting optimal treatment plans for new patients (out-of-sample) and typical data



(a) PICP (b) $\overline{|C|}$

Figure 3: IHDP: the results obtained under the in-sample setting.



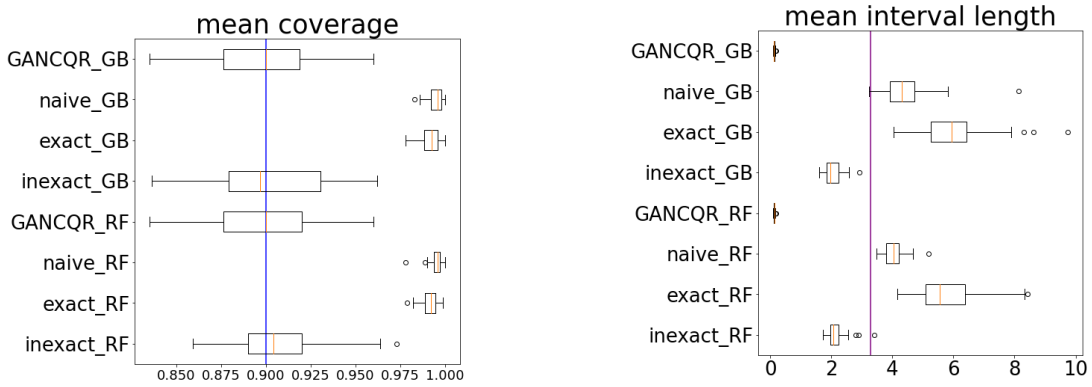
(a) PICP (b) $\overline{|C|}$

Figure 4: Synthetic data: the results obtained under the out-of-sample setting.

analysis (in-sample). Notably, GANCQR outperforms state-of-the-art methods, particularly the exact method, which generates significantly wider prediction intervals in the in-sample setting compared to the intervals in the out-of-sample setting.

5 CONCLUSIONS

In this paper, we introduced GANCQR for predictive inference of ITE on datasets afflicted by two types of selection bias: self-selection bias and attrition selection bias. Through numerical demonstrations, we showcased that GANCQR exhibits favorable performance in terms of both validity and efficiency when compared to state-of-the-art methods. The slight undercoverage of GANCQR may be due to the small sample size and estimation errors incurred by the GANITE component. To further enhance performance, we can improve GANITE to reduce estimation errors and leverage the generative capabilities of GANs to increase the sample size.



(a) PICP (b) $\overline{|C|}$

Figure 5: Synthetic data: the results obtained under the in-sample setting.

REFERENCES

Angelopoulos, A. N. and S. Bates. 2023. "Conformal Prediction: A Gentle Introduction". *Foundations and Trends® in Machine Learning* 16(4):494–591.

Athey, S., J. Tibshirani, and S. Wager. 2019. "Generalized Random Forests". *The Annals of Statistics* 47(2):1148–1178.

Chernozhukov, V., K. Wüthrich, and Y. Zhu. 2021. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls". *Journal of the American Statistical Association* 116(536):1849–1864.

Cochran, W. G. and D. B. Rubin. 1973. "Controlling Bias in Observational Studies: A Review". *Sankhyā: The Indian Journal of Statistics, Series A* 35(4):417–446.

Curth, A., R. W. Peck, E. McKinney, J. Weatherall and M. van Der Schaar. 2024. "Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities". *Clinical Pharmacology and Therapeutics* 115(4):710–719.

Diaconis, P. and D. Freedman. 1980. "Finite Exchangeable Sequences". *The Annals of Probability* 8(4):745–764.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil. 2008. "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects". *Econometrica* 76(5):1191–1206.

Fontana, M., G. Zeni, and S. Vantini. 2023. "Conformal Prediction: A Unified Review of Theory and New Challenges". *Bernoulli* 29(1):1–23.

Foster, J. C., J. M. Taylor, and S. J. Ruberg. 2011. "Subgroup Identification from Randomized Clinical Trial Data". *Statistics in Medicine* 30(24):2867–2880.

Ge, Q., X. Huang, S. Fang, S. Guo, Y. Liu, W. Lin *et al*. 2020. "Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection". *Frontiers in Genetics* 11:585804.

Ghosheh, G. O., J. Li, and T. Zhu. 2024. "A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records". *ACM Computing Surveys* 56(6):1–34.

Grimmer, J., S. Messing, and S. J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods". *Political Analysis* 25(4):413–434.

Haynes, L., B. Goldacre, and D. J. Torgerson. 2012. "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials". *Available at SSRN: 2131581*.

Heckman, J. J., H. Ichimura, and P. E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *The Review of Economic Studies* 64(4):605–654.

Heckman, J. J. and R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions: An Overview". *Journal of Econometrics* 30(1):239–267.

Hill, J. L. 2011. "Bayesian Nonparametric Modeling for Causal Inference". *Journal of Computational and Graphical Statistics* 20(1):217–240.

Imbens, G. W. and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An introduction.* United Kingdom: Cambridge University Press.

Kuang, K., Y. Li, B. Li, P. Cui, H. Yang, J. Tao *et al*. 2021. "Continuous Treatment Effect Estimation via Generative Adversarial De-confounding". *Data Mining and Knowledge Discovery* 35:2467–2497.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning". *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Lei, L. and E. J. Candès. 2021, 10. "Conformal Inference of Counterfactuals and Individual Treatment Effects". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83(5):911–938.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel and M. Welling. 2017. "Causal Effect Inference with Deep Latent-Variable Models". In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Volume 30, 6449–6459. New York: Curran Associates Inc.

McDermott, M. B. A., T. Yan, T. Naumann, N. Hunt, H. Suresh, P. Szolovits *et al*. 2018. "Semi-Supervised Biomedical Translation with Cycle Wasserstein Regression GANs". In *Proceedings of the AAAI Conference on Artificial Intelligence*, edited by M. Wooldridge, J. Dy, and S. Natarajan, Volume 32, 2363–2370. Washington, DC: Association for the Advancement of Artificial Intelligence Press.

Mendil, M., L. Mossina, and D. Vigouroux. 2023. "PUNCC: a Python Library for Predictive Uncertainty Calibration and Conformalization". In *Proceedings of Machine Learning Research*, edited by H. Papadopoulos, K. A. Nguyen, H. Boström, and L. Carlsson, 1–20. New York: Curran Associates Inc.

Mirza, M. and S. Osindero. 2014. "Conditional Generative Adversarial Nets". *arXiv preprint arXiv:1411.1784*.

Nie, X. and S. Wager. 2021. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects". *Biometrika* 108(2):299–319.

Prosperi, M., Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He *et al*. 2020. "Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare". *Nature Machine Intelligence* 2:369–375.

Romano, Y., E. Patterson, and E. Candes. 2019. "Conformalized Quantile Regression". In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Volume 32, 3543–3553. New York: Curran Associates Inc.

Rubin, D. B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions". *Journal of the American Statistical Association* 100(469):322–331.

Shafer, G. and V. Vovk. 2008. "A Tutorial on Conformal Prediction". *Journal of Machine Learning Research* 9:371–421.

Shalit, U., F. D. Johansson, and D. Sontag. 2017. "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms". In *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh, Volume 70, 3076–3085. New York: Curran Associates, Inc.

Van der Laan, M. J. and S. Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

Vovk, V., A. Gammerman, and G. Shafer. 2005. *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag.

Xiao, C., E. Choi, and J. Sun. 2018. "Opportunities and Challenges in Developing Deep Learning Models Using Electronic Health Records Data: A Systematic Review". *Journal of the American Medical Informatics Association* 25(10):1419–1428.

Yoon, J., J. Jordon, and M. Schaar. 2018. "GAIN: Missing Data Imputation Using Generative Adversarial Nets". In *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, Volume 80, 5689–5698. New York: Curran Associates, Inc.

Yoon, J., J. Jordon, and M. van der Schaar. 2018. "GANITE: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets". In *Proceedings of the 6th International Conference on Learning Representations*, edited by Y. Bengio and Y. LeCun, Volume 3, 2196–2217. New York: Curran Associates, Inc.

AUTHOR BIOGRAPHIES

JIAXING WANG is a Ph.D. student in the Operations Research program at North Carolina State University. Her email address is jwang97@ncsu.edu and her website is https://www.or.ncsu.edu/people/jwang97/.

HONG WAN is an Associate Professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. Her research interests focus on the fields of complex simulation modeling, data analysis, and experimental design. Her email address is hwan4@ncsu.edu and her website is https://www.ise.ncsu.edu/people/hwan4/.

XI CHEN is an Associate Professor in the Grado Department of Industrial and Systems Engineering at Virginia Tech. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization. Her email address is xchen.ise@vt.edu and her website is https://sites.google.com/vt.edu/xi-chen-ise/home.