# Impossibility of Distribution-Free Predictive Inference for Individual Treatment Effects

**Chongguang Tao*** **Zheng Zhou*** **Yuhong Yang†**
Qiuzhen College and Yau Mathematical Sciences Center, Tsinghua University
{taocg21, zhouz23}@mails.tsinghua.edu.cn, yyangsc@tsinghua.edu.cn

## Abstract

Uncertainty quantification for individual treatment effects (ITEs) is a daunting challenge in causal inference. Motivated by recent advances in conformal prediction, several works aim to construct distribution-free prediction sets for ITEs with desired coverage under standard assumptions such as strong ignorability and overlap. In this paper, we show that such goals are fundamentally unattainable. Specifically, we establish finite-sample and asymptotic impossibility results demonstrating that any distribution-free prediction set achieving desired coverage for ITEs must be trivial, in the sense that it has infinite expected length. This infinite-expected-length conclusion requires no regularity assumption on the covariates and holds even in their absence. Our analysis builds on ideas from the hardness of conditional independence testing and exploits the missing-data nature of causal inference. These results provide a new perspective on existing methods, clarifying that their apparent success necessarily relies on additional structural assumptions beyond standard causal assumptions.

## 1 Introduction

Quantifying uncertainty for individual treatment effects (ITEs) is a daunting challenge in modern causal inference, with applications ranging from personalized medicine to policy evaluation [1–3]. Although substantial progress has been made in estimating conditional average treatment effects (CATEs) [4–8], uncertainty quantification for ITEs remains considerably more difficult and is especially appealing in high-stakes settings, where decisions are made for individuals rather than populations. A natural goal, therefore, is to construct prediction sets for ITEs with valid coverage.

Conformal prediction provides a general framework for constructing prediction sets with finite-sample, distribution-free coverage guarantees [9–12], and has recently been extended to causal inference [13–16]. In particular, conformal methods for counterfactual inference [13] and conformal meta-learners [15] suggest that nontrivial predictive inference for ITEs may be achievable in some settings. This naturally raises an essential question: under standard causal assumptions such as strong ignorability and overlap in treatment assignment [17], is distribution-free predictive inference for ITEs actually possible?

In this paper, we answer this question in the negative. Under strong ignorability and standard overlap conditions, we establish finite-sample and asymptotic impossibility results showing that distribution-free prediction sets with valid or asymptotically valid coverage for the ITE must be uninformative. With continuous covariates, such sets must cover each pre-specified real value with high probability, yielding infinite expected length in finite samples and diverging expected length asymptotically. We further show that the infinite-expected-length conclusion requires no regularity assumption on the covariates and holds even in their absence (Proposition A.1). These results show that informative distribution-free predictive inference for ITEs is infeasible without additional structural assumptions.

*These authors are co-first authors, contributed equally to this work, and are listed in alphabetical order.
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

At a technical level, our analysis builds on ideas from **the hardness of conditional independence testing** [18]. For continuous covariates, we connect valid predictive inference for ITEs to conditional independence testing, a problem that is fundamentally intractable in a distribution-free sense. We further develop a complementary argument that removes the covariate continuity requirement for the infinite-expected-length conclusion. The result also reflects two difficulties specific to causal inference. First, ITE inference inherits the difficulty of **counterfactual prediction under covariate shift**: calibration across treatment groups depends on the propensity score $e(x)$, whose unknown nature makes reliable reweighting intrinsically difficult [13, 19]. This hardness propagates directly to methods such as [13], which construct ITE prediction sets by first predicting counterfactual outcomes. Second, the **missing data nature of causal inference** [20] creates a deeper non-identifiability: for each unit, one potential outcome is unobserved, so different data-generating processes can induce the same observed distribution while implying arbitrarily different ITEs. Our proof exploits this indistinguishability, forcing any uniformly valid prediction set to be uninformative.

Our analysis provides a new perspective on existing methods for ITE inference. The impossibility result implies that any procedure producing nontrivial prediction sets must rely on assumptions beyond strong ignorability and overlap. This clarifies the role of additional conditions in prior work: conformal inference for counterfactuals [13] relies on **knowledge or accurate estimation of the propensity score** to validly reweight across treatment groups, whereas conformal meta-learners [15] depend on **stochastic dominance conditions** that cannot hold uniformly over the model class. Thus, the apparent success of existing methods reflects additional structure rather than a resolution of the fundamental hardness of ITE inference.

**Our contributions.** First, we establish finite-sample and asymptotic impossibility results for distribution-free predictive inference of individual treatment effects under strong ignorability and standard overlap, showing that any uniformly valid procedure must be trivial and have infinite expected length. The infinite-expected-length conclusion requires no regularity assumption on the covariates and holds even in their absence. Second, we explain this impossibility through two fundamental features of the problem, namely the hardness of conditional independence testing and the missing-data nature of causal inference, and also give a positive result under additional structural assumptions, showing that informative ITE prediction is possible under sensible additional conditions in the standard causal framework. Third, we revisit representative conformal methods for ITE inference, showing that their apparent success actually relies on additional assumptions, and use simulations to illustrate their limitations in practice.

## 2 Main Results

### 2.1 Notations and Setup

We work in the potential outcomes framework with binary treatment [20]. For units $i = 1, \ldots, n$, let $T_i \in \{0, 1\}$ denote the treatment indicator, $Y_i(1)$ and $Y_i(0)$ the potential outcomes, and $X_i \in \mathbb{R}^p$ the covariates, with at least one coordinate being continuously distributed. The observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. Under the superpopulation framework, we assume that

$$(X_i, Y_i(1), Y_i(0), T_i) \overset{\text{i.i.d.}}{\sim} P, \qquad i = 1, \ldots, n,$$

where $P$ is an unknown joint distribution on $(X, Y(1), Y(0), T)$. Let $\mathcal{O}_n = \{(X_i, Y_i, T_i)\}_{i=1}^n$ denote the observed data, and let $\overline{\mathcal{O}}_n = \{(X_i, Y_i(1), Y_i(0), T_i)\}_{i=1}^n$ denote the corresponding full data. Our estimand of interest is the individual treatment effect (ITE), $\Delta = Y(1) - Y(0)$.

Let $\mathcal{E}$ denote the class of all probability distributions on $(X, Y(1), Y(0), T)$ such that at least one coordinate of $X$ is absolutely continuous with respect to the one-dimensional Lebesgue measure and $T$ is binary. Define the strong-ignorability class [17] by

$$\mathcal{P}_{\text{SI}} = \big\{P \in \mathcal{E} : (Y(1), Y(0)) \perp T \mid X\big\}.$$

In causal inference problems, one additionally imposes an overlap condition on the propensity score $e_P(X) := P(T = 1 \mid X)$ [17]. We suppress the subscript $P$ (distribution) and write $e(X)$ when there is no ambiguity. In the main text, we focus on the strong-overlap class

$$\mathcal{P}_{\text{SI}}^{\text{str}} := \Big\{P \in \mathcal{P}_{\text{SI}} : e_P(X) \in [c_P, 1 - c_P] \text{ a.s. for some } c_P > 0\Big\}.$$

That is, the propensity score is bounded away from 0 and 1, with the constant $c_P$ allowed to depend on $P$. In Appendix A.2, we also consider broader overlap-based classes.

A (possibly randomized) prediction set for the ITE is a mapping

$$\widehat{C}_{n,\alpha} : \mathbb{R}^p \to \mathcal{B}(\mathbb{R}) := \{\text{Borel subsets of } \mathbb{R}\},$$

constructed from the observed data $\mathcal{O}_n$. In the impossibility results, a randomized procedure may additionally use a seed $U \sim \mathrm{Unif}(0,1)$, independent of all sampled units, with the same seed distribution under every candidate law $P$. We write $\widehat{C}_{n,\alpha}(x;u)$ when displaying this dependence; otherwise, the seed is suppressed, and probabilities and expectations include its randomness. Consider a new unit from the superpopulation with covariates $X$ and the ITE $\Delta$. We say that $\widehat{C}_{n,\alpha}$ achieves finite-sample and asymptotic distribution-free validity at level $1-\alpha$ over the class $\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$ if, respectively,

$$\sup_{P\in\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha, \qquad \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha.$$

In the standard supervised-learning setting, conformal prediction can achieve the above finite-sample distribution-free validity under exchangeability [9–12]. Our results below show that this guarantee does not extend to the causal setting: even under strong ignorability and overlap assumptions, distribution-free validity for ITE prediction is impossible without additional structural assumptions.

### 2.2 Impossibility Results under Overlap Assumptions

In this subsection, we state our main results, which show that without additional model assumptions, nontrivial inference with a coverage guarantee for the ITE $\Delta$ is impossible. For this reason, we refer to these results as **impossibility results**. Here and below, $\mathrm{length}$ denotes Lebesgue measure. We begin with the asymptotic result under strong overlap; the finite-sample counterpart is stated in Appendix A.3. If the prediction set only achieves asymptotic distribution-free validity at level $1-\alpha$ over $\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$, we have the following asymptotic impossibility result:

**Theorem 2.1** (Asymptotic impossibility under strong overlap)**.** Suppose a sequence of prediction sets $\{\widehat{C}_{n,\alpha}\}_{n\ge1}$ satisfies the asymptotic distribution-free validity at level $1-\alpha$ over $\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$:

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha.$$

Then, for any $P \in \mathcal{P}_{\mathrm{SI}}$ and any $y \in \mathbb{R}$,

$$\limsup_{n\to\infty} \mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha,$$

and hence

$$\lim_{n\to\infty} \mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big] = \infty.$$

The asymptotic theorem shows that any prediction set achieving distribution-free validity for $\Delta$ at level $1-\alpha$ must have asymptotic inclusion probability at least $1-\alpha$ for each pre-specified real number, and hence must have diverging expected length. Appendix A.3 gives the corresponding finite-sample version.

*Proof sketch.* Given the full-data sample $\overline{\mathcal{O}}_{n+1}$ and the seed $U$, let $\eta_n(\overline{\mathcal{O}}_{n+1}, U)$ be the indicator that $Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1}; U)$.

- *Remove overlap by perturbing treatment.* For each $n$, replace treatment with an independent fair coin flip with probability $\varepsilon$. Letting $\varepsilon \downarrow 0$ gives $\sup_{P\in\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{E}_{P,U}[\eta_n] = \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\eta_n]$ (Lemma A.3).
- *Remove the ignorability restriction using a continuous covariate.* For each fixed $n$, an absolutely continuous coordinate of $X$ permits partitions into arbitrarily many equal-probability cells. Lemma A.1 uses these cells to construct a mixture of $(n+1)$-sample laws from $\mathcal{P}_{\mathrm{SI}}$ that agrees with $Q^{\otimes(n+1)}$ whenever the sampled covariates fall in distinct cells. The agreement is preserved after adjoining the independent seed. This event becomes arbitrarily likely, so Lemma A.2 and the assumed bound give $\limsup_{n\to\infty} \mathbb{E}_{Q,U}[\eta_n] \le \alpha$ for every fixed $Q \in \mathcal{E}$.

- *Express exclusion of $y$ as miscoverage.* For fixed $P \in \mathcal{P}_{\mathrm{SI}}$ and $y \in \mathbb{R}$, complete the missing potential outcomes to obtain $Q_y \in \mathcal{E}$ with the same observed-data law as $P$ and $\Delta = y$ almost surely. Thus, for every $n$, $\mathbb{P}_P\{y \notin \widehat{C}_{n,\alpha}(X)\} = \mathbb{P}_{Q_y}\{\Delta \notin \widehat{C}_{n,\alpha}(X)\} = \mathbb{E}_{Q_y,U}[\eta_n]$. The bound from the second step gives the pointwise inclusion claim.
- The expected-length conclusion then follows from Tonelli's theorem and Fatou's lemma.

**Remark 2.1.** In Proposition A.1 of Appendix A.2, we prove that infinite expected length holds without any regularity assumption on $X$, even when there are no covariates, provided coverage holds uniformly over the corresponding strong-overlap class. The proof idea is that on the positive-probability event that all $n$ units are control units, we can assign the unobserved potential outcomes so that the ITE equals any prescribed $y$. Uniform coverage and strong overlap then give a positive inclusion-probability bound that holds uniformly in $y$.

### 2.3 A Positive Result under Additional Structural Assumptions

The impossibility results above show that strong ignorability and overlap alone do not suffice for distribution-free predictive inference for the ITE with finite expected length, even in the absence of covariates. A natural way to obtain informative intervals is to impose additional structural assumptions on the data-generating process. We illustrate this in a simple, admittedly restrictive setting, not to claim broad applicability, but to show that informative prediction is possible beyond the minimal causal assumptions. Specifically, we consider an additive model in which the random error components are independent of the covariates and the two potential outcomes are conditionally independent given the covariates.

For $t \in \{0, 1\}$, let

$$\mu_t(x) := \mathbb{E}[Y(t) \mid X = x], \qquad \tau(x) := \mu_1(x) - \mu_0(x), \qquad \varepsilon_t := Y(t) - \mu_t(X).$$

Denote the treated and control indices by

$$\mathcal{T} := \{i : T_i = 1\}, \qquad \mathcal{C} := \{i : T_i = 0\}.$$

**Assumption 2.1.** Suppose that:

(i) The joint distribution of $(X, Y(1), Y(0), T)$ satisfies strong ignorability and strong overlap;

(ii) for $t \in \{0, 1\}$, the residuals and the potential outcomes satisfy: $\varepsilon_t \perp X$, $Y(1) \perp Y(0) \mid X$.

Under Assumption 2.1, let $\xi := \varepsilon_1 - \varepsilon_0$. Then

$$\Delta = Y(1) - Y(0) = \tau(X) + \xi,$$

and $\xi$ is independent of $X$. This suggests a splitting strategy: estimate $\tau(\cdot)$ on one part and recover the distribution of $\xi$ on the other. We formalize this construction in Algorithm 1. Under the assumptions above, the following result establishes the validity of Algorithm 1.

**Algorithm 1** Split-and-pair prediction interval for the ITE

**Require:** Data $\{(X_i, T_i, Y_i)\}_{i=1}^n$, level $\alpha \in (0, 1)$, split ratio $\rho_n \in (0, 1)$.

1: Sample without replacement subsets $\mathcal{T}_1 \subset \mathcal{T}$ and $\mathcal{C}_1 \subset \mathcal{C}$ of sizes $\lfloor \rho_n |\mathcal{T}| \rfloor$ and $\lfloor \rho_n |\mathcal{C}| \rfloor$, respectively; define $\mathcal{I}_1 := \mathcal{T}_1 \cup \mathcal{C}_1$, $\mathcal{T}_2 := \mathcal{T} \setminus \mathcal{T}_1$, and $\mathcal{C}_2 := \mathcal{C} \setminus \mathcal{C}_1$.

2: Using $\mathcal{I}_1$, estimate $\hat{\mu}_1(\cdot)$ and $\hat{\mu}_0(\cdot)$, and set $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$.

3: For $i \in \mathcal{T}_2$ and $j \in \mathcal{C}_2$, define $\hat{\varepsilon}_i^{(1)} := Y_i - \hat{\mu}_1(X_i)$ and $\hat{\varepsilon}_j^{(0)} := Y_j - \hat{\mu}_0(X_j)$.

4: Let $m_n := \min\{|\mathcal{T}_2|, |\mathcal{C}_2|\}$. Randomly pair $m_n$ treated residuals and $m_n$ control residuals without replacement, yielding pairs $(i_1, j_1), \ldots, (i_{m_n}, j_{m_n})$, and form $\hat{W}_k := \hat{\varepsilon}_{i_k}^{(1)} - \hat{\varepsilon}_{j_k}^{(0)}$ for $k = 1, \ldots, m_n$.

5: Let $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ be the empirical $\alpha/2$- and $(1 - \alpha/2)$-quantiles of $\{\hat{W}_k\}_{k=1}^{m_n}$.

6: Output $\hat{C}_{n,\alpha}(x) := \big[\hat{\tau}(x) + \hat{q}_{\alpha/2},\ \hat{\tau}(x) + \hat{q}_{1-\alpha/2}\big]$.

**Theorem 2.2.** Suppose Assumption 2.1 holds. Assume further that:

(i) for each $t \in \{0,1\}$, $\|\hat{\mu}_t - \mu_t\|_\infty = o_P(1)$;

(ii) the distribution of $\xi$ admits a density that is continuous, bounded, and strictly positive in neighborhoods of its $\alpha/2$- and $1-\alpha/2$-quantiles;

(iii) $n\rho_n \to \infty$ and $n(1-\rho_n) \to \infty$ as $n \to \infty$.

Then

$$\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})\Big) = 1 - \alpha + o(1).$$

Theorem 2.2 shows that, under the structural assumptions above, the proposed interval is asymptotically valid without being as conservative as some other procedures, as will be seen in Section 4. Note that the condition in (i) of the theorem is quite mild and is satisfied by standard nonparametric estimators [21, 22]; the conditions in (ii) and (iii) are non-restrictive. The conditional independence in Assumption 2.1 is most essential for our positive results. In Appendix A.5, we further prove a stronger conditional version and show that its coverage error is governed by two terms: the first-stage regression error in estimating the conditional mean functions and the second-stage quantile estimation error in calibrating the paired residual differences. This decomposition highlights the role of the treatment-stratified splitting ratio $\rho_n$, which controls the tradeoff between these two sources of error. We also discuss this tradeoff in more detail under additional smoothness assumptions on $\mu_t$.

# 3 Reviewing Some Previous Methods

Our impossibility results provide a new perspective on previous methods for ITE inference. In particular, they show that nontrivial prediction intervals cannot be obtained purely under strong ignorability and overlap alone. Existing procedures should therefore be understood as depending, explicitly or implicitly, on assumptions that restrict the problem sufficiently to permit informative intervals. Since some of the positive results on ITE inference do not seem to have stated all the conditions or subtle requirements explicitly, they may be interpreted as valid distribution-free predictive inference. We focus on two important results below for a complementary and nuanced understanding of the foundational matter.

## 3.1 Lei and Candès (2021): conformal inference for counterfactuals and ITEs

We begin with [13], which develops a conformal approach to inferring ITEs by treating the missing counterfactual as the primary prediction target. For an in-study unit, one observes only

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0),$$

so only one of $Y_i(1)$ and $Y_i(0)$ is available. The method first constructs an interval $\widehat{C}_t(x)$ for the missing potential outcome $Y(t)$ at covariate value $X = x$ for the group with $T_i = 1 - t$, and then converts these counterfactual intervals into an interval for $\Delta_i = Y_i(1) - Y_i(0)$. This yields a concrete conformal framework for inferring counterfactuals, and hence ITEs, for units with exactly one observed potential outcome. The construction uses weighted conformal inference, with weights determined by the propensity score $e(x) = \mathbb{P}(T = 1 \mid X = x)$. When the propensity score is known, as in completely randomized or stratified randomized experiments, the resulting procedure yields finite-sample marginal coverage for counterfactual outcomes, and therefore for the derived ITE intervals, without parametric assumptions on the outcome distribution. In observational studies, where $e(x)$ must be estimated, the paper establishes a doubly robust approximate validity.

However, the problem changes qualitatively when one moves from in-study units to units outside the study: for a new unit, neither potential outcome is observed. In [13], the first response is a naïve Bonferroni construction based on the counterfactual intervals. Using the first-stage intervals $\widehat{C}_1(x)$ and $\widehat{C}_0(x)$, it forms the out-of-study ITE interval

$$\widehat{C}^{\mathrm{Bon}}_{\Delta}(x) := \widehat{C}_1(x) - \widehat{C}_0(x) = \{u - v : u \in \widehat{C}_1(x),\ v \in \widehat{C}_0(x)\}.$$

If each counterfactual interval attains coverage level $1 - \alpha/2$, then a Bonferroni argument gives target coverage $1 - \alpha$ for $\Delta$. But this guarantee already rests on the validity of the first-stage counterfactual

intervals, which in [13] are constructed by weighted conformal inference and therefore require either knowledge of, or sufficiently accurate estimation of, the propensity score $e(x)$. In observational studies, estimating $e(x)$ can itself be difficult, especially with high-dimensional covariates [23, 24]. Even when the first-stage coverage is reliable, the direct Bonferroni combination is often highly conservative. To mitigate this, the authors propose nested approaches, including an inexact version that regresses the endpoints of in-study ITE intervals on the covariates and then applies the fitted maps to a new covariate value $x$. As the authors note, however, this method remains inexact and does not provide a formal coverage guarantee for the ITE of a new unit outside the study. The following numerical example illustrates that there actually can be a systematic discrepancy in coverage regardless of the sample size.

**Example 3.1.** Suppose

$$Y(1), Y(0) \overset{\text{ind}}{\sim} N(0,1), \qquad T \sim \text{Bernoulli}(1/2), \qquad X \perp (Y(1), Y(0), T).$$

For an in-study unit, conditional on the observed potential outcome, the ITE has variance 1, so the exact $90\%$ interval uses the $N(0,1)$ quantile $z_{0.95}$. Regressing the endpoints of these exact in-study intervals on $X$ therefore yields, since $X$ is uninformative, the constant interval $[-z_{0.95}, z_{0.95}]$ for a new unit. But for a new unit outside the study, $\Delta = Y(1) - Y(0) \sim N(0,2)$, so this interval has coverage only $2\Phi(z_{0.95}/\sqrt{2}) - 1 \approx 0.755$, well below $0.9$.

In this sense, while the inexact method attempts to extend the first-stage construction, it may provide systematically unreliable results for ITE prediction outside the study. To address this lack of a guarantee, the authors further propose an exact nested method for out-of-study ITE prediction. The method applies a second conformalization step to the induced interval-valued data $(X_i, \widehat{C}_i)$, where $\widehat{C}_i$ is the first-stage ITE interval for an in-study unit, thereby turning the nested construction from a heuristic transfer rule into a procedure with a rigorous coverage guarantee for a new unit.

More specifically, letting $C$ denote the interval-valued target for a new unit, the exact nested method intends to construct an interval-valued map $\widehat{\mathcal{C}}(x)$ such that

$$\mathbb{P}\big(C \subset \widehat{\mathcal{C}}(X)\big) \geq 1 - \gamma.$$

Combined with the first-stage guarantee, this would yield the guaranteed coverage for the ITE of a unit outside the study. The validity of the above exact nested method does not follow from overlap and strong ignorability alone: it also depends on sufficiently accurate estimation of the propensity score, and hence on additional conditions on the propensity score function. Moreover, even when these extra assumptions restore validity, both [13] and our numerical evidence suggest that the method remains conservative, sometimes producing intervals longer than the naive Bonferroni construction. More broadly, this highlights the central difficulty of out-of-study ITE prediction: achieving validity without excessive conservativeness generally requires assumptions beyond overlap and strong ignorability alone.

### 3.2 Conformal meta-learners for ITE inference

We briefly review the conformal meta-learner framework of [15]. At a high level, it combines a meta-learning stage that constructs an estimator $\hat{\tau}(x)$ from the potential-outcome regression functions, or related nuisance components, using a T-learner, S-learner, X-learner, or another plug-in approach [6], with a conformal calibration stage that turns this estimate into an interval for the ITE at a new covariate value $x$ while remaining robust to misspecification of the nuisance learners.

A key technical ingredient in [15] is the comparison between an *oracle* nonconformity score $V^*$, defined by the unobserved ITE $\Delta$, and an *estimated* (pseudo) score $V_\varphi$ obtained by replacing $\Delta$ with a pseudo-outcome $\tilde{Y}$. The pseudo score is the quantity used for conformal calibration: it preserves the functional form of the oracle score while remaining computable from the observed data, and under suitable stochastic ordering conditions relative to the oracle score, [15] derive a coverage guarantee for the resulting conformal pseudo-intervals $\widehat{C}_\varphi(X_{n+1})$.

**Theorem (Alaa et al., 2023).** Assuming exchangeability of $(X_i, T_i, Y_i(0), Y_i(1))_{i=1}^{n+1}$, there exists a distribution-dependent threshold $\alpha^* \in (0,1)$ such that the pseudo-interval $\widehat{C}_\varphi(X_{n+1})$ satisfies

$$\mathbb{P}\Big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}_\varphi(X_{n+1})\Big) \geq 1 - \alpha, \qquad \forall \alpha \in (0, \alpha^*),$$

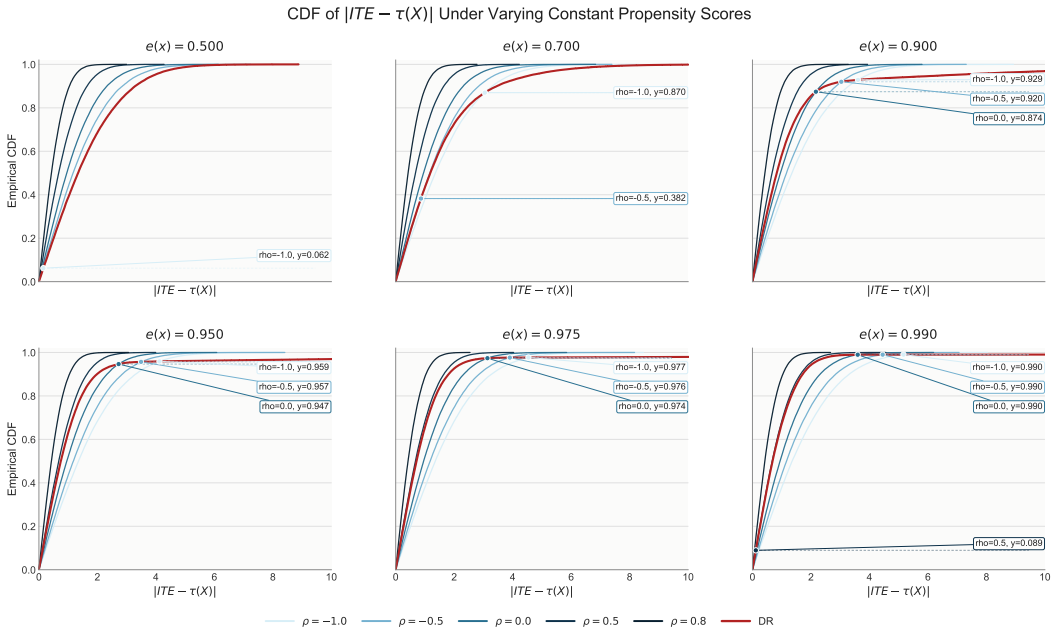



Figure 1: CDFs of the DR pseudo score and the true score in a simple setting with known regression functions and known constant propensity scores. The intersections vary substantially across distributions, and the effective validity threshold $\alpha^*$ can be arbitrarily close to 0.

provided that at least one of the following stochastic ordering conditions holds: (i) $V_\varphi \succeq_{(1)} V^*$, (ii) $V_\varphi \preceq_{(2)} V^*$, and (iii) $V_\varphi \succeq_{\mathrm{mcx}} V^*$. Under condition (i), we have $\alpha^* = 1$.

For completeness, the formal definitions of these three order relations are collected in Appendix B. Indeed, were the dominance to hold globally, it would yield nontrivial prediction intervals with distribution-free validity at any pre-specified level $\alpha$, contradicting the hardness results of Section 2.

**Corollary 3.1.** Suppose that, for all distributions $P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$, the pseudo score $V_\varphi$ satisfies the first-order stochastic dominance $V_\varphi \succeq_{(1)} V^*$. Then the conformal pseudo-interval $\widehat{C}_\varphi$ satisfies

$$\mathbb{P}_P\big(y \notin \widehat{C}_\varphi(X)\big) \le \alpha, \quad \forall P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}},\ \forall y \in \mathbb{R}.$$

Consequently, any such procedure must produce prediction sets with infinite expected length, i.e.,

$$\sup_{P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_\varphi(X)\big)\Big] = \infty.$$

This observation bears directly on distribution-dependent validity thresholds, denoted by $\alpha^*$. For a fixed data-generating distribution $P$, one may define a maximal level $\alpha^*(P)$ below which coverage holds, but our results show that $\alpha^*(P)$ cannot be bounded uniformly away from zero over the model class; in particular, there exists a sequence of distributions $\{P_n\}$ such that $\alpha^*(P_n) \to 0$.

**Corollary 3.2.** Suppose that, for all distributions $P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$, the distribution-dependent validity thresholds $\alpha^*(P)$ satisfy

$$\alpha^*(P) \ge \epsilon > 0.$$

Then for $0 < \alpha \le \epsilon$, the conformal pseudo-interval $\widehat{C}_\varphi$ satisfies

$$\mathbb{P}_P\big(y \notin \widehat{C}_\varphi(X)\big) \le \alpha, \quad \forall P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}},\ \forall y \in \mathbb{R}.$$

Consequently, any such procedure must produce prediction sets with infinite expected length, i.e.,

$$\sup_{P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}} \mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_\varphi(X)\big)\Big] = \infty.$$

From a practical perspective, this sharply limits the usefulness of guarantees of the form "valid for all $\alpha \leq \alpha^*$." In applications, users must choose a nominal level $\alpha$ without knowing the distribution-dependent threshold $\alpha^*(P)$; our simulations show that even in seemingly benign settings, this effective validity threshold may be arbitrarily close to 0.

To illustrate this point, Figure 1 considers a simple setting with known regression functions and propensity score, and compares the cdf of the DR pseudo score with that of the corresponding true score across different values of the dependence parameter $\rho$ and different constant propensity scores. Even in this favorable setting, the pseudo score does not uniformly dominate the true score: the two cdfs intersect substantially across distributions and, in some cases, only at very small tail probabilities. Equivalently, the admissible target level, or validity threshold $\alpha^*$, can be arbitrarily small and may approach 0. This is fully consistent with Corollary 3.1 and Corollary 3.2: the target level is inherently distribution-dependent, so no nontrivial $\alpha$ can be chosen uniformly over the whole model class.

## 4 Experiments

In this section, we conduct simulation studies to illustrate the practical implications of our theoretical results. We compare several representative methods for predictive inference on individual treatment effects (ITEs), with a focus on empirical coverage and interval length.

Our simulations are based on the design of [13], which in turn is adapted from [5], with a modification tailored to ITE inference. Let $X = (X_1, X_2)^\top$, where

$$X_j = \Phi(X'_j), \qquad j = 1, 2,$$

$\Phi$ is the standard normal cdf, and $(X'_1, X'_2)$ is a bivariate Gaussian vector with mean zero and marginal variance one. We consider both independent and correlated covariates, as in [13].

The treated potential outcome is generated from

$$\mathbb{E}[Y(1) \mid X] = f(X_1)f(X_2), \qquad f(x) = \frac{2}{1 + \exp\{-12(x - 0.5)\}},$$

and we set

$$Y(1) = f(X_1)f(X_2) + \sigma(X)\varepsilon_1, \qquad Y(0) = \sigma(X)\varepsilon_0.$$

We consider

$$\mathrm{Corr}(\varepsilon_0, \varepsilon_1) \in \{-1, -0.5, 0\},$$

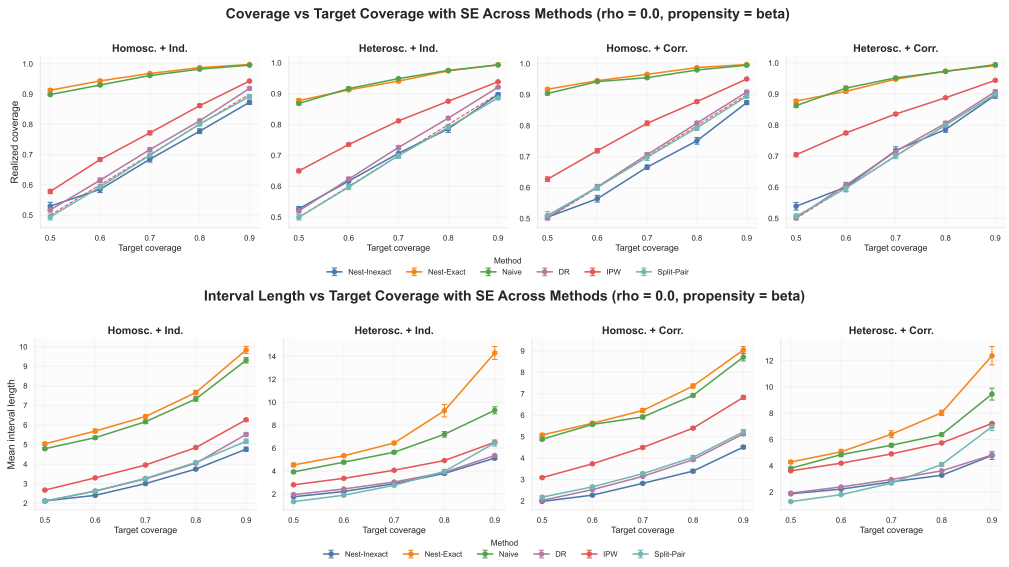



Figure 2: Top: realized coverage versus nominal target coverage under the beta propensity score setting with independent errors $\varepsilon_0$ and $\varepsilon_1$. The red dashed line is the identity line. Curves below it fail to attain the target coverage. Bottom: average interval length versus nominal target coverage under the same setting.

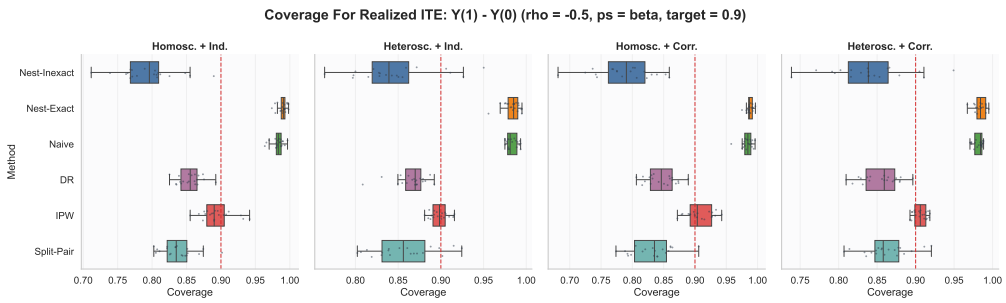



Figure 3: Empirical coverage of realized ITE intervals under the beta propensity score setting with $\mathrm{Corr}(\varepsilon_0, \varepsilon_1) = -0.5$. The red dashed line marks the nominal coverage level 0.9.

to allow $\varepsilon_0$ and $\varepsilon_1$ to be dependent in the first two correlation cases. We focus on negative correlations because they yield a more challenging setting for ITE prediction.

For treatment assignment, we consider two propensity score regimes. In the covariate-dependent setting, following [13], we set

$$e(X) = \mathbb{P}(T = 1 \mid X) = \frac{1}{4}\big(1 + \beta_{2,4}(X_1)\big),$$

where $\beta_{2,4}$ is the cdf of the $\mathrm{Beta}(2, 4)$ distribution. In the constant-propensity setting, we take

$$e(X) = p, \qquad p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}.$$

Together, these regimes let us examine performance under both covariate-dependent treatment assignment and varying degrees of treatment imbalance. In the implementations where the propensity score is treated as unknown, we estimate $e(X)$ using a boosting-based learner rather than plugging in the oracle propensity score. Additional implementation details are provided in Appendix C.1.

Unlike [13], we do not set $Y(0) = 0$. This is because ITE inference is fundamentally different from CATE estimation or counterfactual prediction: one cannot in general assume that the control potential outcome is non-random or independent of $Y(1)$. Allowing $Y(0)$ to remain random and correlated with $Y(1)$ therefore gives a more realistic setting for inferring $\Delta = Y(1) - Y(0)$.

We begin with the beta propensity score setting $e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$ and independent errors $\varepsilon_0$ and $\varepsilon_1$. Figure 2 reports realized coverage, together with the corresponding average interval length, for six methods: Naive, Nest-Exact, and Nest-Inexact from [13]; DR and IPW from [15]; and our Split-Pair method. Several features stand out. Nest-Inexact does not consistently attain the nominal target level, and this failure becomes more pronounced in later settings. Although Nest-Exact enjoys a formal guarantee, it is highly conservative in practice, often even more so than the Naive method. DR and IPW appear somewhat better calibrated in this particular setting, but they remain generally conservative and, as the later experiments show, this behavior is not uniform across settings. By contrast, Split-Pair attains the target coverage level across all four scenarios in this setting.

We next consider the same beta propensity score setting but allow dependence between the two potential outcomes. Figure 3 reports the empirical coverage when $\mathrm{Corr}(\varepsilon_0, \varepsilon_1) = -0.5$. In this setting, Nest-Inexact no longer attains the target level 0.9, and this agrees with the example given earlier. We also see that Split-Pair fails to attain the target in this case. This is expected, since $Y(1)$ and $Y(0)$ are now correlated, so the setting no longer satisfies the applicability condition of our method.

Figure 4 considers a more challenging setting with $\mathrm{Corr}(\varepsilon_0, \varepsilon_1) = -1$ and constant propensity score $e(X) = 0.9$, and plots realized coverage against the nominal target level. In this regime, Nest-Inexact, DR, and IPW all fail to track the target level closely, with realized coverage remaining well below the diagonal over a wide range of targets. This is consistent with Corollary 3.1 and Corollary 3.2: for conformal meta-learners, the nominal target level cannot be chosen uniformly over the model class, since validity necessarily depends on the underlying distribution and no single nontrivial $\alpha$ yields a distribution-free guarantee across all data-generating processes. Thus, even when DR and IPW appear reasonably calibrated in some settings, this behavior cannot persist uniformly.

We also study a substantially harder stress-test design, described in detail in Appendix C.2. In this design, the covariate space is partitioned into a fine checkerboard, and the propensity score alternates sharply between high- and low-probability cells. This makes estimation of $e(X)$ by default learners

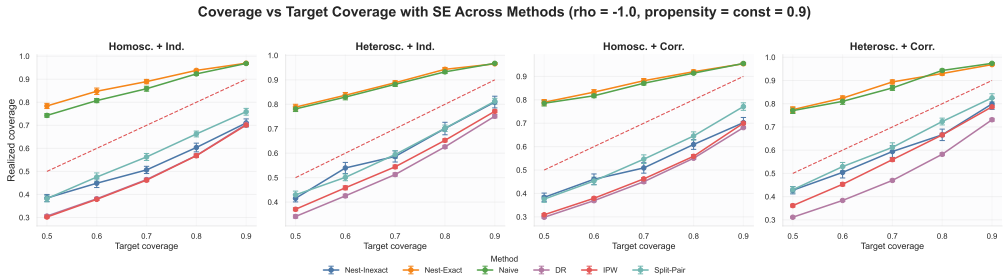



Figure 4: Realized coverage versus nominal target coverage when $\mathrm{Corr}(\varepsilon_0, \varepsilon_1) = -1$ and $e(X) = 0.9$. The red dashed line is the identity line. Curves below it fail to attain the target coverage.

much more difficult in finite samples. Figure 5 shows that in this regime all methods fall materially below the identity line across the target levels we consider. Most notably, the conservative behavior of Naive and Nest-Exact seen in the earlier experiments no longer persists: once the propensity structure becomes sufficiently difficult, even these methods fail to attain the nominal target coverage. This also highlights the difficulty of distribution-free predictive inference for individual treatment effects.

## 5 Conclusions

Distribution-free predictive inference has achieved tremendous success in classical regression and related contexts. In this paper, we studied the feasibility of distribution-free predictive inference for ITEs under the standard causal assumptions of strong ignorability and overlap. Our main conclusion is unfortunately negative: any prediction set that attains valid distribution-free coverage for the ITE must be trivial, in the sense of having infinite expected length, and this impossibility holds both in finite samples and asymptotically. This infinite-expected-length conclusion requires no regularity assumption on the covariates and holds even in their absence. Our analysis provides a fresh perspective on existing methods by clarifying that informative conformal procedures for ITE inference must rely on additional structure beyond the standard causal framework. Under additional assumptions, we also obtain a positive result by proposing the Split-Pair method, which yields informative predictive inference for ITEs. Our simulation studies reinforce this point, showing that methods designed to produce shorter intervals may suffer from undercoverage, whereas methods with stronger formal guarantees often become markedly conservative. Taken together, these results show that the difficulty of distribution-free ITE inference is intrinsic to the missing-data nature of the problem.

## Acknowledgements and Disclosures

The authors thank the Yau Mathematical Sciences Center and Qiuzhen College at Tsinghua University for supporting this research and their participation in the conference. The authors declare no competing interests.

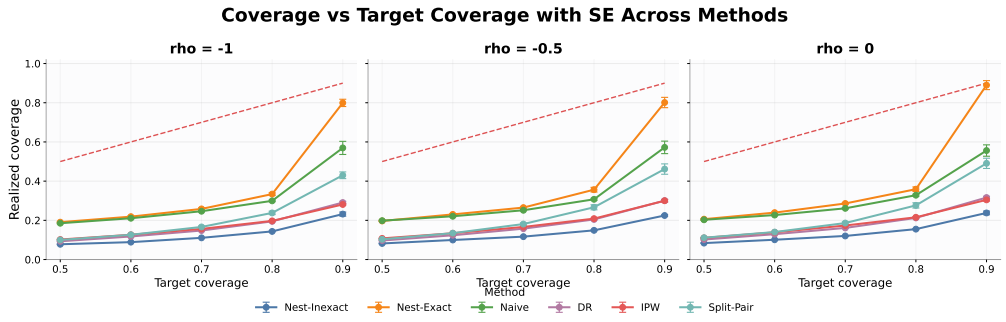



Figure 5: Realized coverage versus nominal target coverage in the checkerboard stress-test design. The three panels correspond to $\rho \in \{-1, -0.5, 0\}$. The red dashed line is the identity line.

## References


[1] David M Kent, Ewout Steyerberg, and David Van Klaveren. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Bmj*, 363, 2018.

[2] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[3] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

[4] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[5] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[6] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[7] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[8] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[9] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

[10] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[11] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL `https://doi.org/10.1561/2200000101`.

[12] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2026. URL `https://arxiv.org/abs/2411.11824`.

[13] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83 (5):911–938, 2021.

[14] Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

[15] Ahmed M Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. *Advances in neural information processing systems*, 36:47682–47703, 2023.

[16] Baozhen Wang and Xingye Qiao. Conformal inference of individual treatment effects using conditional density estimates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21009–21016, 2025.

[17] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[18] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

[19] Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.

[20] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[21] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[22] Xiaohong Chen and Timothy M Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.

[23] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, 2021.

[24] James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

# A Proof of Main Theorems

We provide the proofs of the main theorems in Section 2 in the Appendix. The Appendix is organized as follows:

- A.1. we prove a technical lemma and state its asymptotic consequence;
- A.2. we state the impossibility results over $\mathcal{P}_{\mathrm{SI}}$, which serves as the foundation for the proofs of Theorems A.3 and 2.1;
- A.3. we present the proofs of Theorems A.3 and 2.1 based on the results in A.2;
- A.4. we present the proofs of the results stated in A.2;
- A.5. we present the proof of Theorem 2.2 and further discussion on the splitting ratio.

## A.1 Technical Lemmas

Let $\mathcal{D} := \mathbb{R}^p \times \mathbb{R}^2 \times \{0,1\}$ denote the full-data sample space. A possibly randomized test is a measurable map

$$\psi_n : \mathcal{D}^n \times [0,1] \to \{0,1\}.$$

Here $U$ is independent of the sample and all auxiliary variables. Writing $\lambda$ for the law of $U$, we use $\mathbb{E}_{P,U}$ for expectation under $P^{\otimes n} \otimes \lambda$.

**Lemma A.1.** Consider any (possibly randomized) test

$$\psi_n : \mathcal{D}^n \times [0,1] \to \{0,1\}$$

based on the full data $\overline{\mathcal{O}}_n = \{(X_i, Y_i(1), Y_i(0), T_i)\}_{i=1}^n$. If

$$\sup_{P \in \mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\psi_n(\overline{\mathcal{O}}_n, U)] \le \alpha \quad \text{for some } \alpha \in (0,1),$$

then it necessarily holds that

$$\sup_{Q \in \mathcal{E}} \mathbb{E}_{Q,U}[\psi_n(\overline{\mathcal{O}}_n, U)] \le \alpha.$$

*Proof.* Fix $Q \in \mathcal{E}$ and choose a coordinate $X^{(r)}$ with an absolutely continuous marginal law under $Q$. For any positive integer $k$, its continuous distribution function partitions the covariate space into $k$ Borel cells $A_1, \dots, A_k$, each with $Q_X$-probability $1/k$. Let $b(x)$ be the cell index. Write $q(x) = Q(T = 1 \mid X = x)$, $q_1(x) = q(x)$, and $q_0(x) = 1 - q(x)$. Since the sample space is standard Borel, choose regular conditional kernels $K_t(x, \cdot) = Q((Y(1), Y(0)) \in \cdot \mid X = x, T = t)$; their values where $q_t(x) = 0$ are arbitrary.

For $v = (v_1, \dots, v_k) \in (0,1)^k$, define $g_v(x) = \mathbf{1}\{v_{b(x)} < q(x)\}$. Construct $P_v$ by drawing $X \sim Q_X$, setting $T = g_v(X)$, and drawing $(Y(1), Y(0)) \sim K_{g_v(X)}(X, \cdot)$ conditionally on $X$. Then $P_v \in \mathcal{P}_{\mathrm{SI}}$.

Define the probability measure $M_k$ on $\mathcal{D}^n$ by $M_k(B) = \int_{(0,1)^k} P_v^{\otimes n}(B)\, dv$ for each measurable $B \subseteq \mathcal{D}^n$. We have $\mathbb{E}_{M_k \otimes \lambda}[\psi_n] = \int_{(0,1)^k} \mathbb{E}_{P_v^{\otimes n} \otimes \lambda}[\psi_n]\, dv \le \alpha$.

Let $D_k = \{b(X_i) \ne b(X_j) \text{ for all } 1 \le i < j \le n\}$. For $x_1, \dots, x_n$ in distinct cells and $t_1, \dots, t_n \in \{0,1\}$,

$$\int_{(0,1)^k} \prod_{i=1}^n \mathbf{1}\{g_v(x_i) = t_i\}\, dv = \prod_{i=1}^n q_{t_i}(x_i).$$

It follows that $M_k(B \cap D_k) = Q^{\otimes n}(B \cap D_k)$ for every measurable $B \subseteq \mathcal{D}^n$. Consequently, $M_k \otimes \lambda$ and $Q^{\otimes n} \otimes \lambda$ agree on $D_k \times [0,1]$.

By the union bound, $Q^{\otimes n}(D_k^c) \le \binom{n}{2}/k$, since $Q^{\otimes n}\{b(X_i) = b(X_j)\} = \sum_{\ell=1}^k Q_X(A_\ell)^2 = 1/k$ for each $i < j$. Hence, using $0 \le \psi_n \le 1$,

$$\begin{aligned}\mathbb{E}_{Q^{\otimes n} \otimes \lambda}[\psi_n] &= \mathbb{E}_{M_k \otimes \lambda}[\psi_n \mathbf{1}_{D_k}] + \mathbb{E}_{Q^{\otimes n} \otimes \lambda}[\psi_n \mathbf{1}_{D_k^c}] \\ &\le \mathbb{E}_{M_k \otimes \lambda}[\psi_n] + Q^{\otimes n}(D_k^c) \le \alpha + \frac{\binom{n}{2}}{k}.\end{aligned}$$

Letting $k \to \infty$ gives $\mathbb{E}_{Q,U}[\psi_n(\overline{\mathcal{O}}_n, U)] \le \alpha$. Taking the supremum over $Q$ proves the claim. □

The following asymptotic statement is a direct consequence of Lemma A.1.

**Lemma A.2.** Let $\{\psi_n\}_{n\geq 1}$ be a sequence of (possibly randomized) tests with $\psi_n : \mathcal{D}^n \times [0,1] \to \{0,1\}$. Then

$$\sup_{Q\in\mathcal{E}} \limsup_{n\to\infty} \mathbb{E}_{Q,U}[\psi_n(\overline{\mathcal{O}}_n, U)] \;\leq\; \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\psi_n(\overline{\mathcal{O}}_n, U)].$$

### A.2 Impossibility Results over overlap sub-classes of $\mathcal{P}_{\mathrm{SI}}$

In this appendix, we also consider the following overlap-based sub-classes of $\mathcal{P}_{\mathrm{SI}}$:

$$\mathcal{P}_{\mathrm{SI}}^{\mathrm{ov}} := \Big\{P \in \mathcal{P}_{\mathrm{SI}} : 0 < e_P(X) < 1 \text{ a.s.}\Big\},$$

$$\mathcal{P}_{\mathrm{SI}}^{(r)} := \Big\{P \in \mathcal{P}_{\mathrm{SI}} : \mathbb{E}_P\big[e_P(X)^{-r}(1-e_P(X))^{-r}\big] < +\infty\Big\}, \quad r \in (0,+\infty),$$

$$\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}} := \Big\{P \in \mathcal{P}_{\mathrm{SI}} : e_P(X) \in [c_P, 1-c_P] \text{ a.s. for some } c_P > 0\Big\}.$$

Here $\mathcal{P}_{\mathrm{SI}}^{\mathrm{ov}}$ is the weakest overlap condition. The class $\mathcal{P}_{\mathrm{SI}}^{(r)}$ requires a finite inverse moment of $e(X)(1-e(X))$; in particular, the assumptions used in [13] for counterfactual prediction are equivalent to $P \in \mathcal{P}_{\mathrm{SI}}^{(1)}$. Finally, $\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$ is the usual strong-overlap condition, with the constant $c_P$ allowed to depend on $P$. These classes are nested: for any $0 < r_1 < r_2 < +\infty$,

$$\mathcal{P}_{\mathrm{SI}}^{\mathrm{str}} \subset \mathcal{P}_{\mathrm{SI}}^{(r_2)} \subset \mathcal{P}_{\mathrm{SI}}^{(r_1)} \subset \mathcal{P}_{\mathrm{SI}}^{\mathrm{ov}}.$$

Throughout this subsection, we fix any $\mathcal{P} \in \{\mathcal{P}_{\mathrm{SI}}^{\mathrm{ov}}, \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}, \mathcal{P}_{\mathrm{SI}}^{(r)}\}$, and state the impossibility results over $\mathcal{P}$; their proofs are deferred to Section A.4.

Throughout this subsection, let $\mathcal{O}_n^{\mathrm{obs}} := \{(X_i, T_i, Y_i)\}_{i=1}^n$ denote the observed data, where $Y_i = T_iY_i(1) + (1-T_i)Y_i(0)$. We assume that $(\mathcal{O}_n^{\mathrm{obs}}, x, u, y) \mapsto \mathbf{1}\{y \in \widehat{C}_{n,\alpha}(x;u)\}$ is jointly measurable, and write $\mathrm{length}(A)$ for the Lebesgue measure of a measurable set $A \subseteq \mathbb{R}$. As in the main text, the seed is suppressed in prediction-set notation unless needed, and all probabilities and expectations include its independent uniform law.

**Theorem A.1** (Finite-sample impossibility)**.** Suppose a prediction set $\widehat{C}_{n,\alpha}$ satisfies the uniform finite-sample coverage guarantee

$$\sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \leq \alpha$$

for some $\alpha \in (0,1)$. Then, for any $P \in \mathcal{P}_{\mathrm{SI}}$ and any $y \in \mathbb{R}$,

$$\mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \leq \alpha.$$

Consequently,

$$\mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big] = \infty.$$

Theorem A.1 shows that any prediction set achieving uniform finite-sample validity under ignorability assumptions must, with probability at least $1-\alpha$, contain each pre-specified real number. As a result, such sets are necessarily uninformative, having infinite expected length.

**Theorem A.2** (Asymptotic impossibility)**.** Suppose a sequence of prediction sets $\{\widehat{C}_{n,\alpha}\}_{n\geq 1}$ satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \leq \alpha.$$

Then, for any $P \in \mathcal{P}_{\mathrm{SI}}$ and any $y \in \mathbb{R}$,

$$\limsup_{n\to\infty} \mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \leq \alpha,$$

and hence

$$\lim_{n\to\infty} \mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big] = \infty.$$

Theorem A.2 establishes that even asymptotic distribution-free validity under ignorability assumptions forces prediction sets to be uninformative. The impossibility therefore reflects the hardness of prediction for ITE under the distribution-free assumption.

**Remark A.1.** Our results apply to procedures whose prediction sets depend only on the observed data, the new unit's covariates, and independent auxiliary randomization. In randomized experiments, the known propensity score can be used as an additional input, as in Lei and Candès [13] and the IPW and DR methods of Alaa et al. [15]. Our impossibility results do not directly apply to those procedures when they use this additional input.

Let $\widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}$ be the following class, with no restriction on the marginal law of $X$:

$$\widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}} := \left\{ P \text{ on } (X, Y(1), Y(0), T) : \begin{array}{l} X \in \mathbb{R}^p, \quad Y(1), Y(0) \in \mathbb{R}, \quad T \in \{0,1\}, \\ (Y(1), Y(0)) \perp T \mid X, \\ e_P(X) \in [c_P, 1-c_P] \quad P\text{-a.s. for some } c_P > 0 \end{array} \right\}.$$

**Proposition A.1** (Infinite expected length without covariate regularity)**.** Fix $n \geq 1$ and $\alpha \in (0,1)$. If

$$\sup_{P \in \widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}} P\{\Delta \notin \widehat{C}_{n,\alpha}(X)\} \leq \alpha,$$

then $\mathbb{E}_P[\mathrm{length}(\widehat{C}_{n,\alpha}(X))] = \infty$ for every $P \in \widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}$.

*Proof.* Fix $P \in \widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}$. For each $y \in \mathbb{R}$, define $Q_y$ by

$$\mathcal{L}_{Q_y}(X, Y(0)) = \mathcal{L}_P(X, Y(0)), \qquad Q_y\{Y(1) = Y(0) + y,\ T = 0\} = 1.$$

For $\varepsilon \in (0, 1/2)$, define $Q_{y,\varepsilon}$ by

$$\mathcal{L}_{Q_{y,\varepsilon}}(X, Y(1), Y(0)) = \mathcal{L}_{Q_y}(X, Y(1), Y(0)), \qquad Q_{y,\varepsilon}(T = 1 \mid X, Y(1), Y(0)) = \varepsilon.$$

Then $Q_{y,\varepsilon} \in \widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}$ and $\Delta = y$ almost surely, so

$$1 - \alpha \leq Q_{y,\varepsilon}\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} \leq (1-\varepsilon)^n Q_y\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} + 1 - (1-\varepsilon)^n.$$

Letting $\varepsilon \downarrow 0$ gives $Q_y\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha$.

By strong ignorability, for every $y \in \mathbb{R}$,

$$\begin{aligned} P\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} &\geq P\{T_1 = \cdots = T_n = 0,\ y \in \widehat{C}_{n,\alpha}(X_{n+1})\} \\ &= \mathbb{E}_{Q_y,U}\left[ \mathbf{1}\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} \prod_{i=1}^{n} \{1 - e_P(X_i)\} \right] \\ &\geq c_P^n Q_y\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\} \\ &\geq c_P^n (1 - \alpha). \end{aligned}$$

The lower bound is uniform in $y$. Tonelli's theorem therefore gives

$$\mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X_{n+1})\big)\Big] = \int_{\mathbb{R}} P\{y \in \widehat{C}_{n,\alpha}(X_{n+1})\}\, dy = \infty.$$

□

**Remark A.2.** The conclusion of Proposition A.1 also holds when $X$ takes values in any measurable space $\mathcal{X}$ and its marginal law is restricted to any specified family, provided coverage holds uniformly over the corresponding strong-overlap class and membership in the prediction set is jointly measurable in the observed data, covariates, auxiliary seed, and candidate effect. This includes many standard examples: Gaussian marginals, categorical or text-valued covariates, and a singleton covariate space etc. Indeed, $Q_y$ and $Q_{y,\varepsilon}$ preserve the marginal law of $X$, so the proof applies with $\mathbb{R}^p$ replaced by $\mathcal{X}$.

If $\limsup_{n\to\infty} \sup_{P\in\widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}} P\{\Delta \notin \widehat{C}_{n,\alpha}(X)\} \le \alpha < 1$, then the expected length is infinite for every $P \in \widetilde{\mathcal{P}}_{\mathrm{SI}}^{\mathrm{str}}$ for all sufficiently large $n$. The same eventual conclusion holds for the classes described in the preceding remark.

For each training sample size $n$, the following proofs use the randomized test $\eta_n : \mathcal{D}^{n+1} \times [0,1] \to \{0,1\}$ defined by

$$\eta_n(\overline{\mathcal{O}}_{n+1}, u) = \mathbf{1}\Big\{ Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1}; u) \Big\}.$$

Thus, $\mathbb{E}_{P,U}[\eta_n(\overline{\mathcal{O}}_{n+1}, U)]$ is the miscoverage probability, averaging over the full-data sample and the independent procedure seed.

### A.3 Proof of Theorem A.3 and 2.1

The proofs of the main theorems proceed by verifying the conditions of Theorem A.1 or Theorem A.2. To this end, we first establish the following technical lemma.

**Lemma A.3.** Let

$$\psi_n : \mathcal{D}^n \times [0,1] \to \{0,1\}$$

be any measurable (possibly randomized) test. Then

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{P,U}[\psi_n] = \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\psi_n].$$

**Theorem A.3** (Finite impossibility over overlap sub-classes)**.** Suppose a prediction set $\widehat{C}_{n,\alpha}$ satisfies finite-sample distribution-free validity at level $1-\alpha$ over $\mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha$$

for some $\alpha \in (0,1)$. Then, for any $P \in \mathcal{P}_{\mathrm{SI}}$ and any $y \in \mathbb{R}$,

$$\mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha.$$

Consequently,

$$\mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big] = \infty.$$

Before proving Lemma A.3, we show how it yields the stated theorems; the proof of the lemma is given subsequently.

*Proof of Theorem A.3.* It suffices to verify the condition in Theorem A.1. Applying Lemma A.3 with sample size $n+1$ to the test function $\eta_n$ yields

$$\begin{aligned}
\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) &= \sup_{P\in\mathcal{P}} \mathbb{E}_{P,U}[\eta_n(\overline{\mathcal{O}}_{n+1}, U)] \\
&= \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\eta_n(\overline{\mathcal{O}}_{n+1}, U)] \\
&= \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big).
\end{aligned}$$

Since the left-hand side is at most $\alpha$ by assumption, the desired condition follows. □

*Proof of Theorem 2.1.* It suffices to verify the condition in Theorem A.2. For each $n$, applying Lemma A.3 with sample size $n+1$ to the test function $\eta_n$ gives

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big) = \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{P}_P\Big(\Delta \notin \widehat{C}_{n,\alpha}(X)\Big).$$

Taking the $\limsup$ as $n \to \infty$ and using the assumed asymptotic bound over $\mathcal{P}$ yields the desired condition. □

We now turn to the proof of the lemma.

*Proof of Lemma A.3.* The inequality

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{P,U}[\psi_n] \le \sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\psi_n]$$

is immediate from $\mathcal{P} \subset \mathcal{P}_{\mathrm{SI}}$. It remains to prove the reverse inequality.

Fix any $P_0 \in \mathcal{P}_{\mathrm{SI}}$. Let

$$\mathcal{O} = (X, Y(1), Y(0), T) \sim P_0,$$

and let $B \sim \mathrm{Bernoulli}(1/2)$ be independent of $\mathcal{O}$. For $\varepsilon \in (0,1)$, define an auxiliary Bernoulli variable $A_\varepsilon \sim \mathrm{Bernoulli}(\varepsilon)$, independent of $(\mathcal{O}, B)$, and set

$$T_\varepsilon = (1 - A_\varepsilon)T + A_\varepsilon B.$$

Let $P_{0,\varepsilon}$ denote the law of

$$\mathcal{O}_\varepsilon := (X, Y(1), Y(0), T_\varepsilon).$$

We first verify that $P_{0,\varepsilon} \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}} \subset \mathcal{P}$. Since $(Y(1), Y(0)) \perp T \mid X$ under $P_0$, and $A_\varepsilon, B$ are independent of $(X, Y(1), Y(0), T)$, it follows that

$$(Y(1), Y(0)) \perp T_\varepsilon \mid X.$$

Hence $P_{0,\varepsilon} \in \mathcal{P}_{\mathrm{SI}}$.

Next, writing $e_0(X) = \mathbb{P}(T = 1 \mid X)$ under $P_0$, we have

$$\mathbb{P}(T_\varepsilon = 1 \mid X) = (1-\varepsilon)e_0(X) + \frac{\varepsilon}{2}, \qquad \mathbb{P}(T_\varepsilon = 0 \mid X) = (1-\varepsilon)(1-e_0(X)) + \frac{\varepsilon}{2}.$$

Therefore

$$\min\big(\mathbb{P}(T_\varepsilon = 1 \mid X), \mathbb{P}(T_\varepsilon = 0 \mid X)\big) \ge \frac{\varepsilon}{2} \qquad \text{a.s.},$$

which proves $P_{0,\varepsilon} \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}} \subset \mathcal{P}$.

Since

$$\|P_{0,\varepsilon} - P_0\|_{\mathrm{TV}} \le \mathbb{P}(T_\varepsilon \ne T) \le \varepsilon,$$

we have

$$P_{0,\varepsilon}^{\otimes n} \otimes \lambda \to P_0^{\otimes n} \otimes \lambda \qquad \text{in total variation as } \varepsilon \downarrow 0.$$

Because $0 \le \psi_n \le 1$, it follows that

$$\mathbb{E}_{P_0,U}[\psi_n] = \lim_{\varepsilon\downarrow 0} \mathbb{E}_{P_{0,\varepsilon},U}[\psi_n] \le \sup_{\widetilde{P}\in\mathcal{P}} \mathbb{E}_{\widetilde{P},U}[\psi_n].$$

Taking the supremum over $P_0 \in \mathcal{P}_{\mathrm{SI}}$ yields

$$\sup_{P\in\mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}[\psi_n] \le \sup_{\widetilde{P}\in\mathcal{P}} \mathbb{E}_{\widetilde{P},U}[\psi_n].$$

Combining this with the easy direction completes the proof. □

We next give the proofs of the two auxiliary theorems A.1 and A.2.

### A.4 Proof of Theorems A.1 and A.2

*Proof of Theorem A.1.* Under any $P \in \mathcal{P}_{\mathrm{SI}}$,

$$\mathbb{E}_{P,U}\big[\eta_n(\overline{\mathcal{O}}_{n+1}, U)\big] = \mathbb{P}_P\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) \le \alpha,$$

so $\eta_n$ is a valid level-$\alpha$ test for the corresponding conditional independence null. By Lemma A.1, it follows that for any $Q \in \mathcal{E}$,

$$\mathbb{E}_{Q,U}\big[\eta_n(\overline{\mathcal{O}}_{n+1}, U)\big] \le \alpha, \quad \text{i.e.} \quad \mathbb{P}_Q\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) \le \alpha.$$

Fix any $P \in \mathcal{P}_{\mathrm{SI}}$ and write

$$P = P_{Y(1),Y(0)|X,T} \otimes P_{X,T}.$$

For an arbitrary $y \in \mathbb{R}$, define a distribution $Q \in \mathcal{E}$ on the same space by:

1. The marginal of $(X, T)$ is preserved: $Q_{X,T} = P_{X,T}$.
2. The conditional distribution of $(Y(1), Y(0))$ given $(X, T)$ is

$$Q_{Y(1),Y(0)|X,T=t} = \begin{cases} \mathcal{L}\big(Y(0)+y,\ Y(0)\big),\ \ Y(0) \sim P_{Y(0)|X,T=0}, & t=0, \\ \mathcal{L}\big(Y(1),\ Y(1)-y\big),\ \ Y(1) \sim P_{Y(1)|X,T=1}, & t=1, \end{cases}$$

where $\mathcal{L}(\cdot)$ denotes the law of a random vector.

Under this construction,

$$Q_{(X,T,Y)} = P_{(X,T,Y)},$$

and, moreover,

$$Q_{Y(1)|X,T=1} = P_{Y(1)|X,T=1}, \qquad Q_{Y(0)|X,T=0} = P_{Y(0)|X,T=0},$$

and

$$Q_{Y(1)-Y(0)|X,T=1} = Q_{Y(1)-Y(0)|X,T=0} = \delta_y,$$

so $Q \in \mathcal{E}$ and $\Delta = Y(1) - Y(0) \equiv y$ almost surely under $Q$.

Since the samples are i.i.d. under both $P$ and $Q$, and $Q_{(X,T,Y)} = P_{(X,T,Y)}$ together with $Q_X = P_X$, we have

$$\mathcal{L}_Q\{(X_i, T_i, Y_i)_{i=1}^n, X_{n+1}, U\} = \mathcal{L}_P\{(X_i, T_i, Y_i)_{i=1}^n, X_{n+1}, U\}.$$

Therefore,

$$\mathbb{P}_Q\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) = \mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big).$$

Applying the bound $\mathbb{E}_{Q,U}[\eta_n(\overline{\mathcal{O}}_{n+1}, U)] \le \alpha$ yields

$$\mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) \le \alpha.$$

By the equality of observed-data laws above, it follows that

$$\mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) = \mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big).$$

Hence $\mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha$. This proves the first claim.

To prove the infinite expected length claim, define $f_n(y) = \mathbb{P}_P\big(y \in \widehat{C}_{n,\alpha}(X)\big)$. The previous inequality gives $f_n(y) \ge 1 - \alpha$ for every $y \in \mathbb{R}$. Hence

$$\mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big] = \int_{\mathbb{R}} \mathbb{P}_P\Big(y \in \widehat{C}_{n,\alpha}(X)\Big)\, dy = \int_{\mathbb{R}} f_n(y)\, dy \ \ge\ \int_{\mathbb{R}} (1-\alpha)\, dy = +\infty.$$

The proof is complete. □

*Proof of Theorem A.2.* By the assumed asymptotic validity,

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_{\mathrm{SI}}} \mathbb{E}_{P,U}\big[\eta_n(\overline{\mathcal{O}}_{n+1}, U)\big] \\ &\quad = \limsup_{n\to\infty} \sup_{P \in \mathcal{P}_{\mathrm{SI}}} \mathbb{P}_P\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) \le \alpha. \end{aligned}$$

Applying Lemma A.2, with the test sequence reindexed by the full-data sample size $n+1$, we obtain that for any $Q \in \mathcal{E}$,

$$\limsup_{n\to\infty}\ \mathbb{E}_{Q,U}\big[\eta_n(\overline{\mathcal{O}}_{n+1}, U)\big] \le \alpha,$$

$$\text{i.e.} \quad \limsup_{n\to\infty}\ \mathbb{P}_Q\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) \le \alpha.$$

Fix $P \in \mathcal{P}_{\mathrm{SI}}$ and $y \in \mathbb{R}$, and construct $Q \in \mathcal{E}$ exactly as in the proof of Theorem A.1, so that $\Delta \equiv y$ a.s. under $Q$. Then for each $n$,

$$\mathbb{P}_Q\Big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) = \mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big).$$

Moreover, the same construction preserves the full observed-data law, so for each $n$,

$$\mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big) = \mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X_{n+1})\Big).$$

Therefore,

$$\limsup_{n\to\infty}\ \mathbb{P}_P\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) = \limsup_{n\to\infty}\ \mathbb{P}_Q\Big(y \notin \widehat{C}_{n,\alpha}(X)\Big) \le \alpha.$$

For the divergence of expected length, set $f_n(y) = \mathbb{P}_P\big(y \in \widehat{C}_{n,\alpha}(X)\big)$. Then $\liminf_{n\to\infty} f_n(y) \ge 1-\alpha$ for every $y \in \mathbb{R}$. By Tonelli's theorem, including expectation over the training sample and any auxiliary randomization, $\mathbb{E}_P[\mathrm{length}(\widehat{C}_{n,\alpha}(X))] = \int_{\mathbb{R}} f_n(y)\,dy$. Fatou's lemma therefore gives

$$+\infty = \int_{\mathbb{R}} \liminf_{n\to\infty} f_n(y)\,dy \le \liminf_{n\to\infty} \int_{\mathbb{R}} f_n(y)\,dy = \liminf_{n\to\infty} \mathbb{E}_P\Big[\mathrm{length}\big(\widehat{C}_{n,\alpha}(X)\big)\Big].$$

This proves the stated divergence. □

### A.5 Proof of Theorem 2.2

We prove the following stronger conditional statement, which is a slightly stronger version of Theorem 2.2.

**Proposition A.2.** Suppose Assumption 2.1 holds. Let the treatment-stratified split ratio satisfy $\rho_n \in (0,1)$, $n\rho_n \to \infty$, and $n(1-\rho_n) \to \infty$. Assume moreover that, for each $t \in \{0,1\}$, $\|\hat{\mu}_t - \mu_t\|_\infty = O_P(r_{n,t})$ for some deterministic sequence $r_{n,t} \to 0$, and that the distribution of $\xi := \varepsilon_1 - \varepsilon_0$ admits a density $f_\xi$ which is continuous and strictly positive in neighborhoods of its $\alpha/2$- and $(1-\alpha/2)$-quantiles $q_\ell := F_\xi^{-1}(\alpha/2)$ and $q_u := F_\xi^{-1}(1-\alpha/2)$. Let $m_n := \min\{|\mathcal{T}_2|, |\mathcal{C}_2|\}$, and let

$$\mathcal{D}_n := \sigma\Big(\{(X_i, T_i, Y_i)\}_{i=1}^n,\ \mathcal{T}_1, \mathcal{C}_1,\ (i_1, j_1), \ldots, (i_{m_n}, j_{m_n})\Big)$$

denote the sigma-field generated by the observed sample, the treatment-stratified split, and the random pairing in the second stage of Algorithm 1. Then

$$\Big|\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \,\Big|\, \mathcal{D}_n\Big) - (1-\alpha)\Big| = O_P\Big(r_{n,0} + r_{n,1} + m_n^{-1/2}\Big).$$

*Proof.* Set $\tau(x) := \mu_1(x) - \mu_0(x)$, $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$, and $e_n(x) := \hat{\tau}(x) - \tau(x)$. Also define $a_n := \|\hat{\mu}_1 - \mu_1\|_\infty + \|\hat{\mu}_0 - \mu_0\|_\infty$. Then $\|e_n\|_\infty \le a_n$.

**Step 1: Paired residual differences.** Let $\mathcal{T} := \{i \in \{1,\ldots,n\} : T_i = 1\}$ and $\mathcal{C} := \{i \in \{1,\ldots,n\} : T_i = 0\}$, and denote $N_1 := |\mathcal{T}|$ and $N_0 := |\mathcal{C}|$. After the treatment-stratified split, define $\mathcal{T}_2 := \mathcal{T} \setminus \mathcal{T}_1$ and $\mathcal{C}_2 := \mathcal{C} \setminus \mathcal{C}_1$. For $i \in \mathcal{T}_2$ and $j \in \mathcal{C}_2$, define $\varepsilon_i^{(1)} := Y_i - \mu_1(X_i)$ and $\varepsilon_j^{(0)} := Y_j - \mu_0(X_j)$.

Because $\varepsilon_1 \perp T$, the treated residuals $\{\varepsilon_i^{(1)} : i \in \mathcal{T}_2\}$ are i.i.d. with the same law as $\varepsilon_1$. Likewise, because $\varepsilon_0 \perp T$, the control residuals $\{\varepsilon_j^{(0)} : j \in \mathcal{C}_2\}$ are i.i.d. with the same law as $\varepsilon_0$. Since different units are independent and $\varepsilon_1 \perp \varepsilon_0$, it follows that conditional on $m_n = m$, the random variables $W_k := \varepsilon_{i_k}^{(1)} - \varepsilon_{j_k}^{(0)}$, $k = 1,\ldots,m$, are i.i.d. with common distribution $F_\xi$.

Let $F_m^\circ(x) := \frac{1}{m}\sum_{k=1}^m \mathbf{1}\{W_k \le x\}$ denote their empirical cdf. On the event $\{m_n = m\}$, we write $F_{m_n}^\circ := F_m^\circ$.

**Step 2: The effective second-stage sample size diverges and has order $(1-\rho_n)n$.** Because $P \in \mathcal{P}_{\mathrm{SI}}^{\mathrm{str}}$, there exists $c \in (0, 1/2)$ such that $e(X) = \mathbb{P}(T = 1 \mid X) \in [c, 1-c]$ almost surely. Hence $\pi_1 := \mathbb{P}(T=1) = \mathbb{E}[e(X)] \in [c, 1-c]$ and $\pi_0 := \mathbb{P}(T=0) = 1 - \pi_1 \in [c, 1-c]$. By the law of large numbers, $N_1/n \to \pi_1$ and $N_0/n \to \pi_0$ in probability.

Now $|\mathcal{T}_2| = N_1 - \lfloor \rho_n N_1 \rfloor$ and $|\mathcal{C}_2| = N_0 - \lfloor \rho_n N_0 \rfloor$. Since $n(1-\rho_n) \to \infty$, we obtain $|\mathcal{T}_2|/((1-\rho_n)n) \to \pi_1$ and $|\mathcal{C}_2|/((1-\rho_n)n) \to \pi_0$ in probability. Therefore $m_n/((1-\rho_n)n) \to \min\{\pi_1, \pi_0\}$ in probability. In particular, $m_n \to \infty$ and $m_n^{-1/2} = O_P\big(((1-\rho_n)n)^{-1/2}\big)$.

**Step 3: Quantile estimation error.** Let $\hat{W}_k := \big(Y_{i_k} - \hat{\mu}_1(X_{i_k})\big) - \big(Y_{j_k} - \hat{\mu}_0(X_{j_k})\big)$, $k = 1, \dots, m_n$, and let $\hat{F}_{m_n}(x) := \frac{1}{m_n}\sum_{k=1}^{m_n} \mathbf{1}\{\hat{W}_k \le x\}$ denote the empirical cdf of the plug-in paired differences. Let $\hat{q}_\ell := \hat{F}_{m_n}^{-1}(\alpha/2)$ and $\hat{q}_u := \hat{F}_{m_n}^{-1}(1-\alpha/2)$.

For every $k$, $|\hat{W}_k - W_k| \le \|\hat{\mu}_1 - \mu_1\|_\infty + \|\hat{\mu}_0 - \mu_0\|_\infty = a_n$. Hence, for every $x \in \mathbb{R}$,

$$F^\circ_{m_n}(x - a_n) \le \hat{F}_{m_n}(x) \le F^\circ_{m_n}(x + a_n).$$

Write $\delta_n := \sup_{x\in\mathbb{R}} |F^\circ_{m_n}(x) - F_\xi(x)|$. For any integer $m \ge 1$ and any $u > 0$, conditional on $m_n = m$, the Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F^\circ_m(x) - F_\xi(x)| > u \,\middle|\, m_n = m\right) \le 2e^{-2mu^2}.$$

Therefore, for any $M > 0$, $\mathbb{P}\big(\sqrt{m_n}\,\delta_n > M\big) \le 2e^{-2M^2} + \mathbb{P}(m_n = 0)$. Since $m_n \to \infty$ in probability by Step 2, it follows that $\delta_n = O_P(m_n^{-1/2})$.

Because $f_\xi$ is continuous and strictly positive in neighborhoods of $q_\ell$ and $q_u$, there exist constants $\underline{f} > 0$, $\bar{f} < \infty$, and $\epsilon_0 > 0$ such that $\underline{f} \le f_\xi(x) \le \bar{f}$ for all $x \in [q_\ell - \epsilon_0, q_\ell + \epsilon_0] \cup [q_u - \epsilon_0, q_u + \epsilon_0]$. Consequently, for every $x$ in these neighborhoods, $\left|\hat{F}_{m_n}(x) - F_\xi(x)\right| \le \delta_n + \bar{f} a_n$. Since $F_\xi$ has derivative bounded below by $\underline{f}$ near $q_\ell$ and $q_u$, the standard local quantile perturbation argument yields

$$|\hat{q}_\ell - q_\ell| + |\hat{q}_u - q_u| = O_P(a_n + \delta_n) = O_P\Big(r_{n,0} + r_{n,1} + m_n^{-1/2}\Big).$$

**Step 4: Conditional coverage error.** By construction, $\hat{C}_{n,\alpha}(x) = [\hat{\tau}(x) + \hat{q}_\ell,\ \hat{\tau}(x) + \hat{q}_u]$. For the new point, $\Delta_{n+1} = \tau(X_{n+1}) + \xi_{n+1}$, where $\xi_{n+1} \sim F_\xi$, $\xi_{n+1} \perp X_{n+1}$, and the test point is independent of $\mathcal{D}_n$. Therefore,

$$\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \,\Big|\, \mathcal{D}_n\Big) = \mathbb{E}\big[F_\xi\big(\hat{q}_u + e_n(X_{n+1})\big) - F_\xi\big(\hat{q}_\ell + e_n(X_{n+1})\big) \,\big|\, \mathcal{D}_n\big].$$

On the event $\{|\hat{q}_\ell - q_\ell| \le \epsilon_0/4,\ |\hat{q}_u - q_u| \le \epsilon_0/4,\ \|e_n\|_\infty \le \epsilon_0/4\}$, all arguments of $F_\xi$ above lie inside the neighborhoods on which $f_\xi \le \bar{f}$. Hence, uniformly in $x$,

$$\big|F_\xi\big(\hat{q}_u + e_n(x)\big) - F_\xi(q_u)\big| \le \bar{f}\big(|\hat{q}_u - q_u| + \|e_n\|_\infty\big),$$

and similarly

$$\big|F_\xi\big(\hat{q}_\ell + e_n(x)\big) - F_\xi(q_\ell)\big| \le \bar{f}\big(|\hat{q}_\ell - q_\ell| + \|e_n\|_\infty\big).$$

Since $F_\xi(q_u) - F_\xi(q_\ell) = \big(1 - \frac{\alpha}{2}\big) - \frac{\alpha}{2} = 1 - \alpha$, it follows that

$$\begin{aligned}&\left|\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \,\Big|\, \mathcal{D}_n\Big) - (1-\alpha)\right| \\ &\qquad \le \bar{f}\Big(|\hat{q}_u - q_u| + |\hat{q}_\ell - q_\ell| + 2\|e_n\|_\infty\Big).\end{aligned}$$

Using $\|e_n\|_\infty \le a_n = O_P(r_{n,0} + r_{n,1})$ and the quantile bound established above, we obtain

$$\left|\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \,\Big|\, \mathcal{D}_n\Big) - (1-\alpha)\right| = O_P\Big(r_{n,0} + r_{n,1} + m_n^{-1/2}\Big).$$

This proves Proposition A.2. □

As an immediate consequence, Proposition A.2 implies Theorem 2.2. Let $R_n := \mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \,\Big|\, \mathcal{D}_n\Big) - (1-\alpha)$. Then $R_n = o_P(1)$ by Proposition A.2, and $|R_n| \le 1$. Hence bounded convergence gives $\mathbb{E}[R_n] \to 0$, so $\mathbb{P}_P\Big(\Delta_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})\Big) = 1 - \alpha + o(1)$.

The rate in Proposition A.2 is governed by two terms: the first-stage regression error $r_{n,0} + r_{n,1}$ and the second-stage quantile calibration error $m_n^{-1/2}$. Under strong overlap, $m_n$ is of order $n(1-\rho_n)$ in probability, so increasing $\rho_n$ typically improves first-stage regression estimation but reduces the second-stage calibration sample size. For example, under a nonparametric uniform regression rate of the form $r_{n,t} \asymp (n\rho_n)^{-\beta/(2\beta+d)}$, the choice of $\rho_n$ reflects the balance between $(n\rho_n)^{-\beta/(2\beta+d)}$ and $\{n(1-\rho_n)\}^{-1/2}$. This illustrates that the optimal split depends on the convergence rates of regression estimators and quantile estimators.

## B Definitions of Stochastic Orders

This appendix records the stochastic-order definitions used in our discussion of conformal meta-learners. More specifically, it provides the formal meanings of first-order stochastic dominance, second-order stochastic dominance, and monotone convex order, following the definitions used in [15].

Let $U$ and $V$ be real-valued random variables with cumulative distribution functions $F_U$ and $F_V$.

**Definition B.1** (First-order stochastic dominance)**.** We say that $U$ first-order stochastically dominates $V$, written

$$U \succeq_{(1)} V,$$

if

$$F_U(x) \leq F_V(x), \qquad \forall x \in \mathbb{R},$$

with strict inequality for some $x$.

**Definition B.2** (Second-order stochastic dominance)**.** We say that $U$ second-order stochastically dominates $V$, written

$$U \succeq_{(2)} V,$$

if

$$\int_{-\infty}^{x} \big(F_V(t) - F_U(t)\big)\, dt \geq 0, \qquad \forall x \in \mathbb{R},$$

with strict inequality for some $x$.

**Definition B.3** (Monotone convex order)**.** We say that $U$ dominates $V$ in monotone convex order, written

$$U \succeq_{\mathrm{mcx}} V,$$

if

$$\mathbb{E}[u(U)] \geq \mathbb{E}[u(V)]$$

for every non-decreasing convex function $u : \mathbb{R} \to \mathbb{R}$ for which the expectations are well defined.

In the terminology of [15], first-order stochastic dominance compares location through the ordering of the cdfs, second-order stochastic dominance compares spread through integrated cdfs, and monotone convex order compares expectations against all non-decreasing convex test functions. These are precisely the three order relations invoked in the stochastic-ordering theorem quoted in Section 3.2.

## C Detailed Experimental Setup and Numerical Results

### C.1 Experimental Setup

This appendix gives the full data-generating process and implementation details for the simulation study. The design follows Lei and Candès [13], which is adapted from Wager and Athey [5], with one important modification: both potential outcomes are random, so that the individual treatment effect is itself a random quantity.

**Covariates.** Let $X = (X_1, X_2)^\top$. We generate

$$X_j = \Phi(X'_j), \qquad j = 1, 2,$$

where $\Phi$ is the standard normal cdf. We consider two covariance designs for $(X'_1, X'_2)$. In the independent-covariate design, $(X'_1, X'_2) \sim N(0, I_2)$. In the correlated-covariate design, we generate

$$X'_j = \sqrt{1 - \rho_X}\, Z_j + \sqrt{\rho_X}\, Z_0, \qquad j = 1, 2,$$

where $Z_0, Z_1, Z_2$ are independent standard normal variables and $\rho_X = 0.9$. This transformation produces covariates supported on $[0, 1]^2$, while allowing us to compare independent and strongly correlated covariate settings.

**Potential outcomes.** Define

$$g(x) = \frac{2}{1 + \exp\{-12(x - 0.5)\}}.$$

The conditional mean of the treated potential outcome is

$$\tau(X) = \mathbb{E}[Y(1) \mid X] = g(X_1)g(X_2),$$

and the control mean is zero. The potential outcomes are generated as

$$Y(1) = \tau(X) + \sigma(X)\varepsilon_1, \qquad Y(0) = \sigma(X)\varepsilon_0.$$

We consider two noise-scale settings:

$$\sigma(X) = 1 \quad \text{and} \quad \sigma(X) = -\log(X_1).$$

The first setting is homoscedastic, while the second is heteroscedastic. In the implementation, a small numerical offset is used inside the logarithm to avoid evaluating $\log(0)$.

The pair $(\varepsilon_0, \varepsilon_1)$ is standard normal marginally, with

$$\mathrm{Corr}(\varepsilon_0, \varepsilon_1) \in \{-1, -0.5, 0\}.$$

When the correlation is $-1$, the second error is generated as the corresponding deterministic multiple of the first. Otherwise, we generate

$$\varepsilon_1 = \rho\, \varepsilon_0 + \sqrt{1 - \rho^2}\, \eta, \qquad \eta \sim N(0, 1),$$

independently of $\varepsilon_0$. We focus on non-positive correlations because negative dependence between $Y(0)$ and $Y(1)$ increases the variability of the realized ITE $Y(1) - Y(0)$, making ITE prediction intervals more challenging.

Combining the two covariate designs and the two noise-scale designs gives four scenarios:

Scenario 1: homoscedastic noise, independent covariates,
Scenario 2: heteroscedastic noise, independent covariates,
Scenario 3: homoscedastic noise, correlated covariates,
Scenario 4: heteroscedastic noise, correlated covariates.

**Treatment assignment.** Given $X$, treatment is assigned as

$$T \mid X \sim \mathrm{Bernoulli}(e(X)),$$

and the observed outcome is

$$Y = TY(1) + (1 - T)Y(0).$$

We use two classes of propensity scores. In the covariate-dependent setting, following Lei and Candès [13], we set

$$e(X) = \frac{1}{4}\{1 + \beta_{2,4}(X_1)\},$$

where $\beta_{2,4}$ denotes the cdf of a $\mathrm{Beta}(2, 4)$ distribution. In the constant-propensity setting, we set

$$e(X) = p, \qquad p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}.$$

The constant-propensity settings are included to study the effect of treatment imbalance.

**Difference from the CATE/counterfactual setting.** In the original simulation design of Lei and Candès [13], the focus is on counterfactual prediction and CATE-related quantities. For ITE inference, we do not set $Y(0)$ to be deterministic. Instead, $Y(0)$ remains random and may be correlated with $Y(1)$. This distinction is important because the target of inference is the realized individual treatment effect

$$\tau_{\mathrm{ITE}} = Y(1) - Y(0),$$

not the conditional mean contrast $\mathbb{E}[Y(1) - Y(0) \mid X]$. Allowing both potential outcomes to be random therefore creates a setting closer to the inferential target of individual-level treatment-effect prediction intervals.

**Sample sizes and repetitions.** For each experimental configuration, we generate a training sample of size

$$n_{\text{train}} = 1000$$

and an independent test sample of size

$$n_{\text{test}} = 10000.$$

Each configuration is repeated 20 times using independent random seeds. We evaluate target coverages

$$0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9.$$

**Methods.** We compare the following prediction-interval methods.

- **Nest-Inexact**: the inexact nested conformal ITE method implemented by `cfcausal::conformalIte` with `algo = ‘‘nest’’`, `exact = FALSE`, and CQR.
- **Nest-Exact**: the exact nested conformal ITE method implemented by `cfcausal::conformalIte` with `algo = ‘‘nest’’`, `exact = TRUE`, and CQR.
- **Naive**: the naive ITE conformal method implemented by `cfcausal::conformalIte` with `algo = ‘‘naive’’` and CQR.
- **DR**: a conformal pseudo-outcome meta-learner using the doubly-robust pseudo-outcome.
- **IPW**: a conformal pseudo-outcome meta-learner using the inverse-propensity-weighted pseudo-outcome.
- **Split-Pair**: the split-and-pair ITE method.

For the `cfcausal` methods, we use quantile random forests for the outcome quantile learner and boosting for propensity estimation. For the meta-learners and split-pair method, we use gradient-boosting regressors with the default experiment settings. The split-pair method uses a training/calibration split ratio of $0.5$.

**Evaluation metrics.** On the independent test sample, the true realized ITE

$$Y_i(1) - Y_i(0)$$

is observed by construction. For a fitted interval $[L(X_i), U(X_i)]$, the empirical coverage is

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{L(X_i) \le Y_i(1) - Y_i(0) \le U(X_i)\},$$

and the average interval length is

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \{U(X_i) - L(X_i)\}.$$

The reported tables average these quantities over the 20 Monte Carlo repetitions.

### C.2 Additional Checkerboard Propensity Stress Test

To probe a regime where the propensity score is much harder to learn, we also consider an additional checkerboard stress-test design, implemented in the accompanying code. In this experiment, the covariates are sampled directly from the uniform distribution on the unit square,

$$X \sim \text{Unif}([0,1]^2).$$

We partition $[0,1]^2$ into a $10 \times 10$ grid of equal-sized cells and assign treatment according to an alternating checkerboard pattern. Writing the cell indices as

$$i = \min\{\lfloor 10X_1 \rfloor, 9\}, \qquad j = \min\{\lfloor 10X_2 \rfloor, 9\},$$

we set

$$e(X) = \mathbb{P}(T = 1 \mid X) = \begin{cases} 0.95, & i + j \text{ even}, \\ 0.05, & i + j \text{ odd}. \end{cases}$$

Thus overlap still holds, but the propensity surface is highly discontinuous and oscillates at a much finer scale than in the main experiments.

Figure 6 visualizes this construction. The left and middle panels show the two complementary sets of hard cells: the $Y(0)$-hard cells are exactly those with high propensity, where control units are rare, while the $Y(1)$-hard cells are exactly those with low propensity, where treated units are rare. The right panel shows the corresponding checkerboard propensity surface.

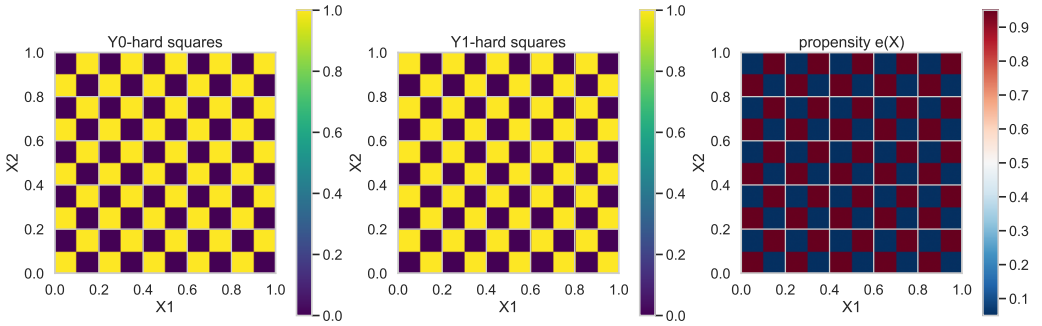



Figure 6: Visualization of the checkerboard stress-test design. Left: cells in which the control potential outcome is harder to learn. Middle: cells in which the treated potential outcome is harder to learn. Right: the checkerboard propensity score $e(X)$, which alternates between $0.95$ and $0.05$ over the $10 \times 10$ grid.

The treatment-effect signal is chosen to be the same as in the main simulation design above. Let $R_0(X)$ be the indicator of the high-propensity cells and $R_1(X) = 1 - R_0(X)$. We then define cell-dependent noise scales

$$\sigma_0(X) = 1 + 9R_0(X), \qquad \sigma_1(X) = 1 + 9R_1(X),$$

so that control outcomes are much noisier in cells where control units are rare and treated outcomes are much noisier in cells where treated units are rare. The potential outcomes are generated as

$$Y(0) = \sigma_0(X)\varepsilon_0, \qquad Y(1) = \tau(X) + \sigma_1(X)\varepsilon_1,$$

where $(\varepsilon_0, \varepsilon_1)$ are standard normal marginally with correlation $\rho \in \{-1, -0.5, 0\}$. The observed outcome is again $Y = TY(1) + (1 - T)Y(0)$.

This construction is deliberately adverse for ITE prediction. In high-propensity cells, there are very few control units precisely where $Y(0)$ is the hardest to learn; in low-propensity cells, there are very few treated units precisely where $Y(1)$ is the hardest to learn. As a result, finite-sample estimation of the propensity score and of the two potential-outcome distributions becomes much more fragile than in the smoother designs above.

We use the same sample sizes, target coverages, and number of Monte Carlo repetitions as in the main experiments, namely $n_{\text{train}} = 1000$, $n_{\text{test}} = 10000$, target coverages $0.5, 0.6, 0.7, 0.8, 0.9$, and 20 repetitions. The set of methods is also the same. For the `cfcausal` methods, we again use CQR with quantile random forests for outcome prediction and boosting for propensity estimation, so these methods must learn the checkerboard propensity from the observed data. For the DR and IPW meta-learners, the implementation follows the accompanying code; the remaining tuning parameters are kept fixed across repetitions. Figure 5 in the main text reports the resulting coverage curves.

Table 1 reports the corresponding mean realized coverage and mean interval length under this checkerboard design.

Table 1: Mean realized coverage and mean interval length under the grid checkerboard DGP. Entries are coverage | length.

| rho | Target | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| -1 | 0.50 | 0.078 \| 2.17 | 0.191 \| 5.33 | 0.185 \| 5.18 | 0.093 \| 2.59 | 0.102 \| 2.85 | 0.100 \| 2.81 |
| | 0.60 | 0.089 \| 2.48 | 0.219 \| 6.13 | 0.211 \| 5.91 | 0.118 \| 3.29 | 0.126 \| 3.55 | 0.126 \| 3.55 |
| | 0.70 | 0.111 \| 3.06 | 0.258 \| 7.29 | 0.246 \| 6.94 | 0.148 \| 4.14 | 0.156 \| 4.37 | 0.166 \| 4.69 |
| | 0.80 | 0.143 \| 4.00 | 0.334 \| 9.53 | 0.300 \| 8.51 | 0.195 \| 5.47 | 0.197 \| 5.56 | 0.238 \| 6.79 |
| | 0.90 | 0.232 \| 6.57 | 0.799 \| 29.45 | 0.570 \| 18.74 | 0.291 \| 8.31 | 0.281 \| 8.01 | 0.430 \| 12.98 |
| -0.5 | 0.50 | 0.082 \| 2.17 | 0.197 \| 5.25 | 0.198 \| 5.27 | 0.097 \| 2.58 | 0.108 \| 2.88 | 0.103 \| 2.77 |
| | 0.60 | 0.100 \| 2.64 | 0.230 \| 6.18 | 0.221 \| 5.92 | 0.124 \| 3.29 | 0.134 \| 3.58 | 0.135 \| 3.61 |
| | 0.70 | 0.117 \| 3.10 | 0.265 \| 7.16 | 0.251 \| 6.77 | 0.157 \| 4.21 | 0.166 \| 4.45 | 0.181 \| 4.90 |
| | 0.80 | 0.149 \| 3.97 | 0.356 \| 9.79 | 0.307 \| 8.37 | 0.205 \| 5.51 | 0.209 \| 5.63 | 0.267 \| 7.36 |
| | 0.90 | 0.225 \| 6.08 | 0.801 \| 28.91 | 0.572 \| 18.02 | 0.301 \| 8.26 | 0.300 \| 8.21 | 0.462 \| 13.64 |
| 0 | 0.50 | 0.084 \| 2.13 | 0.206 \| 5.24 | 0.203 \| 5.18 | 0.102 \| 2.59 | 0.111 \| 2.83 | 0.111 \| 2.86 |
| | 0.60 | 0.101 \| 2.55 | 0.239 \| 6.14 | 0.227 \| 5.79 | 0.129 \| 3.29 | 0.139 \| 3.55 | 0.140 \| 3.60 |
| | 0.70 | 0.120 \| 3.06 | 0.286 \| 7.38 | 0.261 \| 6.72 | 0.161 \| 4.10 | 0.174 \| 4.45 | 0.186 \| 4.84 |
| | 0.80 | 0.155 \| 3.95 | 0.359 \| 9.41 | 0.328 \| 8.55 | 0.212 \| 5.47 | 0.216 \| 5.53 | 0.276 \| 7.31 |
| | 0.90 | 0.238 \| 6.15 | 0.890 \| 37.15 | 0.556 \| 16.55 | 0.316 \| 8.29 | 0.304 \| 7.95 | 0.491 \| 13.94 |

### C.3 Detailed Results for Target Coverage 0.9

We report the detailed numerical results for target coverage $0.9$. The three tables in this subsection correspond to $\rho = 0.0$, $\rho = -0.5$, and $\rho = -1.0$, respectively.

Table 2: Mean realized coverage and mean interval length for target coverage 0.9 across propensity settings when $\rho = 0.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.873 \| 4.76 | 0.998 \| 9.83 | 0.996 \| 9.31 | 0.919 \| 5.52 | 0.943 \| 6.27 | 0.892 \| 5.17 |
| | Heterosc. + Ind. | 0.897 \| 5.13 | 0.994 \| 14.30 | 0.993 \| 9.30 | 0.921 \| 5.36 | 0.939 \| 6.52 | 0.886 \| 6.45 |
| | Homosc. + Corr. | 0.874 \| 4.51 | 0.997 \| 9.04 | 0.995 \| 8.71 | 0.909 \| 5.13 | 0.950 \| 6.83 | 0.895 \| 5.22 |
| | Heterosc. + Corr. | 0.896 \| 4.78 | 0.992 \| 12.38 | 0.995 \| 9.46 | 0.908 \| 4.82 | 0.944 \| 7.22 | 0.899 \| 6.97 |
| const = 0.1 | Homosc. + Ind. | 0.918 \| 6.56 | 0.993 \| 10.02 | 0.995 \| 10.16 | 0.836 \| 4.52 | 0.789 \| 4.57 | 0.864 \| 5.36 |
| | Heterosc. + Ind. | 0.908 \| 8.14 | 0.988 \| 14.14 | 0.994 \| 11.64 | 0.846 \| 4.50 | 0.806 \| 4.89 | 0.872 \| 7.04 |
| | Homosc. + Corr. | 0.901 \| 6.18 | 0.996 \| 10.36 | 0.996 \| 9.57 | 0.829 \| 4.22 | 0.808 \| 5.42 | 0.862 \| 5.15 |
| | Heterosc. + Corr. | 0.906 \| 7.34 | 0.986 \| 13.28 | 0.993 \| 10.14 | 0.830 \| 4.05 | 0.816 \| 5.78 | 0.870 \| 7.26 |
| const = 0.3 | Homosc. + Ind. | 0.872 \| 4.99 | 0.997 \| 9.79 | 0.994 \| 9.14 | 0.913 \| 5.44 | 0.938 \| 6.48 | 0.894 \| 5.24 |
| | Heterosc. + Ind. | 0.887 \| 5.35 | 0.993 \| 14.12 | 0.994 \| 10.02 | 0.918 \| 5.29 | 0.935 \| 6.97 | 0.891 \| 6.75 |
| | Homosc. + Corr. | 0.870 \| 4.54 | 0.998 \| 9.56 | 0.996 \| 8.87 | 0.903 \| 5.05 | 0.941 \| 7.38 | 0.884 \| 5.10 |
| | Heterosc. + Corr. | 0.881 \| 4.65 | 0.992 \| 12.62 | 0.993 \| 8.98 | 0.904 \| 4.83 | 0.937 \| 7.97 | 0.890 \| 6.86 |
| const = 0.5 | Homosc. + Ind. | 0.872 \| 4.70 | 0.998 \| 9.67 | 0.995 \| 9.01 | 0.921 \| 5.51 | 0.944 \| 6.23 | 0.888 \| 5.08 |
| | Heterosc. + Ind. | 0.885 \| 4.86 | 0.991 \| 12.25 | 0.992 \| 9.08 | 0.919 \| 5.26 | 0.939 \| 6.40 | 0.881 \| 6.06 |
| | Homosc. + Corr. | 0.863 \| 4.38 | 0.998 \| 9.18 | 0.996 \| 8.65 | 0.918 \| 5.24 | 0.955 \| 6.95 | 0.903 \| 5.27 |
| | Heterosc. + Corr. | 0.880 \| 4.25 | 0.993 \| 11.92 | 0.993 \| 8.58 | 0.920 \| 4.90 | 0.948 \| 7.30 | 0.888 \| 6.57 |
| const = 0.7 | Homosc. + Ind. | 0.863 \| 4.46 | 0.998 \| 9.35 | 0.996 \| 9.13 | 0.911 \| 5.37 | 0.926 \| 5.83 | 0.887 \| 5.09 |
| | Heterosc. + Ind. | 0.895 \| 4.72 | 0.991 \| 12.71 | 0.992 \| 9.13 | 0.914 \| 5.21 | 0.933 \| 5.89 | 0.875 \| 6.15 |
| | Homosc. + Corr. | 0.829 \| 4.02 | 0.997 \| 9.26 | 0.995 \| 8.65 | 0.912 \| 5.20 | 0.932 \| 6.15 | 0.889 \| 5.16 |
| | Heterosc. + Corr. | 0.871 \| 4.21 | 0.993 \| 12.53 | 0.993 \| 8.49 | 0.917 \| 4.96 | 0.939 \| 6.48 | 0.884 \| 6.64 |
| const = 0.9 | Homosc. + Ind. | 0.888 \| 4.89 | 0.994 \| 8.72 | 0.995 \| 9.34 | 0.850 \| 4.61 | 0.841 \| 4.50 | 0.857 \| 5.01 |
| | Heterosc. + Ind. | 0.887 \| 6.04 | 0.984 \| 12.75 | 0.991 \| 9.93 | 0.862 \| 4.60 | 0.864 \| 4.55 | 0.868 \| 6.38 |
| | Homosc. + Corr. | 0.864 \| 4.63 | 0.995 \| 8.49 | 0.994 \| 8.83 | 0.842 \| 4.36 | 0.851 \| 4.53 | 0.882 \| 5.42 |
| | Heterosc. + Corr. | 0.897 \| 5.41 | 0.987 \| 12.20 | 0.993 \| 11.46 | 0.843 \| 4.03 | 0.884 \| 4.81 | 0.887 \| 7.03 |

Table 3: Mean realized coverage and mean interval length for target coverage 0.9 across propensity settings when $\rho = -0.5$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.794 \| 4.72 | 0.989 \| 9.72 | 0.983 \| 9.16 | 0.855 \| 5.43 | 0.894 \| 6.31 | 0.838 \| 5.25 |
| | Heterosc. + Ind. | 0.847 \| 5.26 | 0.983 \| 13.13 | 0.983 \| 9.85 | 0.865 \| 5.25 | 0.898 \| 6.46 | 0.857 \| 6.59 |
| | Homosc. + Corr. | 0.787 \| 4.44 | 0.988 \| 8.99 | 0.984 \| 8.85 | 0.846 \| 5.17 | 0.907 \| 6.86 | 0.831 \| 5.21 |
| | Heterosc. + Corr. | 0.839 \| 4.62 | 0.984 \| 12.05 | 0.981 \| 8.47 | 0.854 \| 4.82 | 0.906 \| 7.09 | 0.862 \| 6.52 |
| const = 0.1 | Homosc. + Ind. | 0.840 \| 6.05 | 0.975 \| 9.62 | 0.984 \| 10.04 | 0.755 \| 4.47 | 0.724 \| 4.59 | 0.835 \| 5.63 |
| | Heterosc. + Ind. | 0.860 \| 6.98 | 0.976 \| 13.59 | 0.982 \| 11.35 | 0.785 \| 4.42 | 0.761 \| 4.90 | 0.838 \| 7.05 |
| | Homosc. + Corr. | 0.821 \| 5.96 | 0.982 \| 9.99 | 0.988 \| 9.87 | 0.751 \| 4.23 | 0.749 \| 5.42 | 0.820 \| 5.42 |
| | Heterosc. + Corr. | 0.847 \| 6.76 | 0.977 \| 13.35 | 0.987 \| 11.32 | 0.774 \| 4.07 | 0.771 \| 5.72 | 0.827 \| 6.43 |
| const = 0.3 | Homosc. + Ind. | 0.797 \| 4.91 | 0.989 \| 9.99 | 0.981 \| 9.12 | 0.847 \| 5.35 | 0.888 \| 6.50 | 0.821 \| 5.09 |
| | Heterosc. + Ind. | 0.841 \| 5.47 | 0.981 \| 12.61 | 0.981 \| 9.70 | 0.865 \| 5.21 | 0.893 \| 6.86 | 0.861 \| 6.70 |
| | Homosc. + Corr. | 0.771 \| 4.34 | 0.986 \| 9.02 | 0.982 \| 8.77 | 0.835 \| 5.06 | 0.896 \| 7.37 | 0.826 \| 5.19 |
| | Heterosc. + Corr. | 0.826 \| 4.96 | 0.983 \| 12.51 | 0.983 \| 9.20 | 0.852 \| 4.79 | 0.901 \| 7.92 | 0.859 \| 6.66 |
| const = 0.5 | Homosc. + Ind. | 0.792 \| 4.64 | 0.989 \| 9.45 | 0.982 \| 9.13 | 0.857 \| 5.46 | 0.895 \| 6.26 | 0.814 \| 4.99 |
| | Heterosc. + Ind. | 0.827 \| 4.88 | 0.984 \| 13.04 | 0.981 \| 9.30 | 0.870 \| 5.27 | 0.904 \| 6.46 | 0.855 \| 6.19 |
| | Homosc. + Corr. | 0.753 \| 4.13 | 0.988 \| 9.02 | 0.982 \| 8.70 | 0.851 \| 5.19 | 0.915 \| 6.99 | 0.826 \| 5.11 |
| | Heterosc. + Corr. | 0.819 \| 4.36 | 0.986 \| 12.73 | 0.980 \| 8.57 | 0.864 \| 4.85 | 0.912 \| 7.18 | 0.861 \| 6.66 |
| const = 0.7 | Homosc. + Ind. | 0.790 \| 4.49 | 0.991 \| 9.45 | 0.984 \| 9.12 | 0.859 \| 5.53 | 0.874 \| 5.94 | 0.824 \| 5.18 |
| | Heterosc. + Ind. | 0.831 \| 4.51 | 0.984 \| 13.81 | 0.980 \| 9.42 | 0.871 \| 5.25 | 0.894 \| 5.92 | 0.844 \| 6.00 |
| | Homosc. + Corr. | 0.757 \| 4.15 | 0.987 \| 9.14 | 0.981 \| 8.54 | 0.837 \| 5.08 | 0.883 \| 6.18 | 0.824 \| 5.17 |
| | Heterosc. + Corr. | 0.811 \| 4.22 | 0.985 \| 13.43 | 0.981 \| 8.58 | 0.861 \| 4.93 | 0.906 \| 6.57 | 0.858 \| 6.54 |
| const = 0.9 | Homosc. + Ind. | 0.791 \| 4.60 | 0.982 \| 8.66 | 0.980 \| 9.01 | 0.779 \| 4.69 | 0.764 \| 4.53 | 0.803 \| 5.07 |
| | Heterosc. + Ind. | 0.809 \| 4.86 | 0.973 \| 12.34 | 0.981 \| 10.72 | 0.802 \| 4.53 | 0.812 \| 4.62 | 0.834 \| 6.18 |
| | Homosc. + Corr. | 0.792 \| 4.63 | 0.975 \| 8.27 | 0.983 \| 9.01 | 0.758 \| 4.30 | 0.772 \| 4.49 | 0.828 \| 5.41 |
| | Heterosc. + Corr. | 0.835 \| 5.43 | 0.974 \| 12.01 | 0.983 \| 9.95 | 0.775 \| 3.99 | 0.834 \| 4.87 | 0.869 \| 7.58 |

Table 4: Mean realized coverage and mean interval length for target coverage 0.9 across propensity settings when $\rho = -1.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.738 \| 4.71 | 0.977 \| 9.75 | 0.962 \| 9.03 | 0.796 \| 5.38 | 0.845 \| 6.22 | 0.769 \| 5.12 |
| | Heterosc. + Ind. | 0.807 \| 5.23 | 0.975 \| 13.10 | 0.966 \| 9.34 | 0.819 \| 5.23 | 0.869 \| 6.42 | 0.829 \| 6.32 |
| | Homosc. + Corr. | 0.727 \| 4.50 | 0.975 \| 9.33 | 0.954 \| 8.30 | 0.787 \| 5.16 | 0.864 \| 6.83 | 0.776 \| 5.22 |
| | Heterosc. + Corr. | 0.796 \| 6.05 | 0.982 \| 14.11 | 0.970 \| 9.50 | 0.817 \| 4.95 | 0.880 \| 7.27 | 0.849 \| 7.17 |
| const = 0.1 | Homosc. + Ind. | 0.764 \| 5.62 | 0.967 \| 10.24 | 0.972 \| 10.24 | 0.695 \| 4.46 | 0.694 \| 4.83 | 0.769 \| 5.40 |
| | Heterosc. + Ind. | 0.845 \| 7.24 | 0.963 \| 13.63 | 0.978 \| 13.84 | 0.737 \| 4.41 | 0.751 \| 5.20 | 0.798 \| 6.03 |
| | Homosc. + Corr. | 0.770 \| 5.68 | 0.962 \| 9.77 | 0.970 \| 9.35 | 0.682 \| 4.20 | 0.701 \| 5.28 | 0.755 \| 5.23 |
| | Heterosc. + Corr. | 0.817 \| 7.03 | 0.976 \| 14.79 | 0.969 \| 9.92 | 0.726 \| 4.00 | 0.739 \| 5.71 | 0.841 \| 7.30 |
| const = 0.3 | Homosc. + Ind. | 0.745 \| 4.96 | 0.977 \| 10.01 | 0.968 \| 9.45 | 0.796 \| 5.40 | 0.837 \| 6.39 | 0.769 \| 5.12 |
| | Heterosc. + Ind. | 0.793 \| 5.37 | 0.977 \| 15.26 | 0.965 \| 9.33 | 0.815 \| 5.22 | 0.864 \| 6.90 | 0.832 \| 6.49 |
| | Homosc. + Corr. | 0.718 \| 4.45 | 0.976 \| 9.54 | 0.967 \| 8.97 | 0.772 \| 5.01 | 0.866 \| 7.42 | 0.766 \| 5.10 |
| | Heterosc. + Corr. | 0.776 \| 4.67 | 0.977 \| 13.19 | 0.969 \| 9.04 | 0.800 \| 4.79 | 0.871 \| 7.98 | 0.839 \| 6.74 |
| const = 0.5 | Homosc. + Ind. | 0.722 \| 4.56 | 0.976 \| 9.54 | 0.966 \| 9.24 | 0.808 \| 5.52 | 0.845 \| 6.21 | 0.768 \| 5.11 |
| | Heterosc. + Ind. | 0.780 \| 4.82 | 0.971 \| 11.98 | 0.965 \| 9.29 | 0.835 \| 5.34 | 0.870 \| 6.40 | 0.833 \| 6.39 |
| | Homosc. + Corr. | 0.700 \| 4.23 | 0.971 \| 8.98 | 0.964 \| 8.80 | 0.793 \| 5.22 | 0.866 \| 6.82 | 0.775 \| 5.17 |
| | Heterosc. + Corr. | 0.767 \| 4.39 | 0.977 \| 12.34 | 0.960 \| 8.22 | 0.818 \| 4.88 | 0.883 \| 7.30 | 0.842 \| 6.71 |
| const = 0.7 | Homosc. + Ind. | 0.717 \| 4.41 | 0.975 \| 9.42 | 0.961 \| 8.85 | 0.797 \| 5.43 | 0.822 \| 5.91 | 0.768 \| 5.14 |
| | Heterosc. + Ind. | 0.775 \| 4.53 | 0.976 \| 13.75 | 0.965 \| 8.98 | 0.820 \| 5.26 | 0.860 \| 5.95 | 0.826 \| 6.32 |
| | Homosc. + Corr. | 0.707 \| 4.29 | 0.971 \| 9.08 | 0.961 \| 8.65 | 0.772 \| 5.01 | 0.837 \| 6.22 | 0.769 \| 5.14 |
| | Heterosc. + Corr. | 0.792 \| 5.60 | 0.976 \| 11.96 | 0.970 \| 9.57 | 0.805 \| 4.82 | 0.875 \| 6.56 | 0.838 \| 6.77 |
| const = 0.9 | Homosc. + Ind. | 0.710 \| 4.41 | 0.970 \| 9.02 | 0.968 \| 9.33 | 0.704 \| 4.56 | 0.701 \| 4.52 | 0.759 \| 5.18 |
| | Heterosc. + Ind. | 0.809 \| 5.78 | 0.965 \| 12.81 | 0.968 \| 10.90 | 0.751 \| 4.52 | 0.771 \| 4.68 | 0.812 \| 6.38 |
| | Homosc. + Corr. | 0.702 \| 4.37 | 0.954 \| 8.35 | 0.955 \| 8.49 | 0.681 \| 4.15 | 0.700 \| 4.43 | 0.771 \| 5.48 |
| | Heterosc. + Corr. | 0.799 \| 5.44 | 0.968 \| 11.90 | 0.974 \| 10.08 | 0.731 \| 4.04 | 0.787 \| 4.88 | 0.825 \| 7.57 |

### C.4 Detailed Results for Target Coverage 0.8

We report the detailed numerical results for target coverage $0.8$. The three tables in this subsection correspond to $\rho = 0.0$, $\rho = -0.5$, and $\rho = -1.0$, respectively.

Table 5: Mean realized coverage and mean interval length for target coverage 0.8 across propensity settings when $\rho = 0.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.777 \| 3.75 | 0.988 \| 7.66 | 0.982 \| 7.33 | 0.812 \| 4.04 | 0.862 \| 4.85 | 0.801 \| 4.08 |
| | Heterosc. + Ind. | 0.786 \| 3.80 | 0.974 \| 9.27 | 0.975 \| 7.21 | 0.821 \| 3.89 | 0.876 \| 4.92 | 0.791 \| 3.98 |
| | Homosc. + Corr. | 0.751 \| 3.39 | 0.987 \| 7.36 | 0.980 \| 6.92 | 0.808 \| 3.92 | 0.877 \| 5.40 | 0.793 \| 4.03 |
| | Heterosc. + Corr. | 0.786 \| 3.28 | 0.974 \| 8.03 | 0.973 \| 6.38 | 0.806 \| 3.61 | 0.888 \| 5.74 | 0.800 \| 4.10 |
| const = 0.1 | Homosc. + Ind. | 0.838 \| 5.37 | 0.986 \| 9.23 | 0.982 \| 8.24 | 0.703 \| 3.26 | 0.576 \| 3.07 | 0.791 \| 4.38 |
| | Heterosc. + Ind. | 0.804 \| 5.07 | 0.969 \| 10.03 | 0.980 \| 8.91 | 0.704 \| 3.18 | 0.523 \| 2.96 | 0.772 \| 4.16 |
| | Homosc. + Corr. | 0.811 \| 4.73 | 0.990 \| 8.74 | 0.986 \| 7.96 | 0.702 \| 3.17 | 0.554 \| 3.46 | 0.774 \| 4.17 |
| | Heterosc. + Corr. | 0.824 \| 5.44 | 0.973 \| 10.70 | 0.980 \| 7.99 | 0.699 \| 2.92 | 0.500 \| 3.49 | 0.775 \| 4.22 |
| const = 0.3 | Homosc. + Ind. | 0.780 \| 3.99 | 0.987 \| 7.90 | 0.982 \| 7.50 | 0.797 \| 3.91 | 0.849 \| 4.82 | 0.788 \| 4.03 |
| | Heterosc. + Ind. | 0.787 \| 4.14 | 0.970 \| 9.09 | 0.976 \| 7.27 | 0.805 \| 3.75 | 0.869 \| 4.99 | 0.794 \| 4.08 |
| | Homosc. + Corr. | 0.769 \| 3.55 | 0.988 \| 7.58 | 0.983 \| 7.22 | 0.794 \| 3.82 | 0.858 \| 5.55 | 0.793 \| 4.04 |
| | Heterosc. + Corr. | 0.809 \| 3.56 | 0.976 \| 8.64 | 0.977 \| 6.87 | 0.786 \| 3.57 | 0.876 \| 6.13 | 0.790 \| 3.98 |
| const = 0.5 | Homosc. + Ind. | 0.769 \| 3.63 | 0.988 \| 7.67 | 0.983 \| 7.30 | 0.825 \| 4.15 | 0.863 \| 4.79 | 0.780 \| 3.88 |
| | Heterosc. + Ind. | 0.787 \| 3.68 | 0.971 \| 8.62 | 0.974 \| 6.85 | 0.825 \| 3.90 | 0.881 \| 4.96 | 0.790 \| 3.84 |
| | Homosc. + Corr. | 0.751 \| 3.37 | 0.987 \| 7.32 | 0.982 \| 7.10 | 0.819 \| 4.01 | 0.883 \| 5.45 | 0.804 \| 4.09 |
| | Heterosc. + Corr. | 0.796 \| 3.35 | 0.973 \| 8.01 | 0.976 \| 6.56 | 0.822 \| 3.72 | 0.896 \| 5.85 | 0.795 \| 3.95 |
| const = 0.7 | Homosc. + Ind. | 0.768 \| 3.52 | 0.989 \| 7.55 | 0.981 \| 7.05 | 0.795 \| 3.92 | 0.831 \| 4.41 | 0.797 \| 4.08 |
| | Heterosc. + Ind. | 0.798 \| 3.31 | 0.974 \| 9.15 | 0.973 \| 6.55 | 0.804 \| 3.75 | 0.862 \| 4.46 | 0.778 \| 3.71 |
| | Homosc. + Corr. | 0.734 \| 3.24 | 0.985 \| 7.18 | 0.980 \| 6.96 | 0.798 \| 3.84 | 0.855 \| 4.79 | 0.798 \| 4.08 |
| | Heterosc. + Corr. | 0.773 \| 3.19 | 0.975 \| 8.29 | 0.974 \| 6.41 | 0.809 \| 3.61 | 0.881 \| 5.20 | 0.787 \| 3.89 |
| const = 0.9 | Homosc. + Ind. | 0.723 \| 3.30 | 0.987 \| 7.60 | 0.983 \| 7.28 | 0.730 \| 3.41 | 0.714 \| 3.29 | 0.783 \| 4.15 |
| | Heterosc. + Ind. | 0.853 \| 5.01 | 0.974 \| 10.35 | 0.977 \| 7.54 | 0.747 \| 3.30 | 0.760 \| 3.29 | 0.781 \| 4.17 |
| | Homosc. + Corr. | 0.761 \| 3.55 | 0.987 \| 7.50 | 0.985 \| 7.45 | 0.710 \| 3.23 | 0.727 \| 3.31 | 0.789 \| 4.25 |
| | Heterosc. + Corr. | 0.798 \| 3.63 | 0.972 \| 9.24 | 0.981 \| 8.10 | 0.717 \| 2.94 | 0.778 \| 3.31 | 0.807 \| 4.62 |

Table 6: Mean realized coverage and mean interval length for target coverage 0.8 across propensity settings when $\rho = -0.5$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.690 \| 3.74 | 0.967 \| 7.82 | 0.949 \| 7.18 | 0.734 \| 4.07 | 0.785 \| 4.79 | 0.720 \| 4.06 |
| | Heterosc. + Ind. | 0.711 \| 3.63 | 0.952 \| 8.94 | 0.947 \| 7.06 | 0.752 \| 3.87 | 0.823 \| 4.89 | 0.752 \| 3.98 |
| | Homosc. + Corr. | 0.644 \| 3.27 | 0.956 \| 7.21 | 0.950 \| 6.99 | 0.721 \| 3.91 | 0.811 \| 5.35 | 0.714 \| 4.01 |
| | Heterosc. + Corr. | 0.722 \| 3.40 | 0.963 \| 9.19 | 0.949 \| 6.66 | 0.732 \| 3.61 | 0.840 \| 5.70 | 0.750 \| 3.85 |
| const = 0.1 | Homosc. + Ind. | 0.770 \| 5.09 | 0.969 \| 9.05 | 0.961 \| 8.28 | 0.617 \| 3.25 | 0.514 \| 3.07 | 0.748 \| 4.60 |
| | Heterosc. + Ind. | 0.766 \| 5.69 | 0.952 \| 10.76 | 0.948 \| 8.01 | 0.644 \| 3.17 | 0.474 \| 2.98 | 0.750 \| 4.30 |
| | Homosc. + Corr. | 0.719 \| 4.49 | 0.963 \| 8.74 | 0.949 \| 7.28 | 0.624 \| 3.21 | 0.511 \| 3.47 | 0.696 \| 4.14 |
| | Heterosc. + Corr. | 0.756 \| 5.38 | 0.965 \| 13.20 | 0.958 \| 7.65 | 0.632 \| 2.93 | 0.453 \| 3.49 | 0.749 \| 4.27 |
| const = 0.3 | Homosc. + Ind. | 0.678 \| 3.77 | 0.960 \| 7.78 | 0.948 \| 7.34 | 0.720 \| 3.97 | 0.766 \| 4.72 | 0.716 \| 4.04 |
| | Heterosc. + Ind. | 0.711 \| 3.99 | 0.953 \| 9.71 | 0.948 \| 7.27 | 0.743 \| 3.78 | 0.820 \| 5.04 | 0.754 \| 4.11 |
| | Homosc. + Corr. | 0.655 \| 3.38 | 0.964 \| 7.59 | 0.950 \| 7.13 | 0.707 \| 3.81 | 0.805 \| 5.61 | 0.715 \| 4.06 |
| | Heterosc. + Corr. | 0.713 \| 3.47 | 0.954 \| 8.61 | 0.951 \| 6.85 | 0.717 \| 3.57 | 0.829 \| 6.10 | 0.753 \| 4.01 |
| const = 0.5 | Homosc. + Ind. | 0.679 \| 3.61 | 0.968 \| 7.86 | 0.947 \| 7.11 | 0.743 \| 4.14 | 0.792 \| 4.84 | 0.707 \| 3.94 |
| | Heterosc. + Ind. | 0.722 \| 3.68 | 0.947 \| 8.38 | 0.942 \| 6.75 | 0.757 \| 3.90 | 0.827 \| 4.91 | 0.745 \| 3.74 |
| | Homosc. + Corr. | 0.647 \| 3.29 | 0.956 \| 7.15 | 0.951 \| 7.08 | 0.735 \| 4.00 | 0.820 \| 5.45 | 0.722 \| 4.07 |
| | Heterosc. + Corr. | 0.713 \| 3.32 | 0.952 \| 7.93 | 0.947 \| 6.58 | 0.752 \| 3.72 | 0.852 \| 5.84 | 0.758 \| 4.04 |
| const = 0.7 | Homosc. + Ind. | 0.659 \| 3.37 | 0.964 \| 7.51 | 0.951 \| 7.14 | 0.722 \| 3.99 | 0.753 \| 4.43 | 0.707 \| 3.97 |
| | Heterosc. + Ind. | 0.709 \| 3.21 | 0.953 \| 8.71 | 0.944 \| 6.66 | 0.747 \| 3.75 | 0.811 \| 4.52 | 0.745 \| 3.84 |
| | Homosc. + Corr. | 0.656 \| 3.35 | 0.957 \| 7.20 | 0.950 \| 7.02 | 0.710 \| 3.82 | 0.782 \| 4.77 | 0.703 \| 3.94 |
| | Heterosc. + Corr. | 0.698 \| 3.24 | 0.949 \| 7.79 | 0.946 \| 6.61 | 0.731 \| 3.61 | 0.831 \| 5.16 | 0.746 \| 3.88 |
| const = 0.9 | Homosc. + Ind. | 0.683 \| 3.62 | 0.961 \| 7.48 | 0.962 \| 7.60 | 0.632 \| 3.32 | 0.624 \| 3.25 | 0.709 \| 4.14 |
| | Heterosc. + Ind. | 0.750 \| 3.84 | 0.951 \| 10.27 | 0.950 \| 7.63 | 0.672 \| 3.24 | 0.694 \| 3.27 | 0.739 \| 4.00 |
| | Homosc. + Corr. | 0.710 \| 3.83 | 0.963 \| 7.75 | 0.955 \| 7.30 | 0.620 \| 3.20 | 0.635 \| 3.29 | 0.725 \| 4.31 |
| | Heterosc. + Corr. | 0.715 \| 3.65 | 0.955 \| 9.37 | 0.952 \| 7.37 | 0.644 \| 2.95 | 0.712 \| 3.31 | 0.758 \| 4.46 |

Table 7: Mean realized coverage and mean interval length for target coverage 0.8 across propensity settings when $\rho = -1.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.610 \| 3.58 | 0.943 \| 8.00 | 0.909 \| 7.09 | 0.656 \| 3.96 | 0.719 \| 4.71 | 0.639 \| 3.90 |
| | Heterosc. + Ind. | 0.661 \| 3.62 | 0.932 \| 8.64 | 0.924 \| 7.19 | 0.696 \| 3.87 | 0.783 \| 4.89 | 0.715 \| 3.88 |
| | Homosc. + Corr. | 0.598 \| 3.41 | 0.923 \| 7.25 | 0.915 \| 7.11 | 0.659 \| 3.93 | 0.764 \| 5.39 | 0.659 \| 4.05 |
| | Heterosc. + Corr. | 0.651 \| 3.31 | 0.942 \| 8.96 | 0.918 \| 6.59 | 0.678 \| 3.69 | 0.812 \| 5.83 | 0.725 \| 3.99 |
| const = 0.1 | Homosc. + Ind. | 0.690 \| 4.82 | 0.944 \| 9.10 | 0.923 \| 8.08 | 0.561 \| 3.26 | 0.501 \| 3.28 | 0.641 \| 4.11 |
| | Heterosc. + Ind. | 0.747 \| 5.15 | 0.936 \| 10.46 | 0.941 \| 9.15 | 0.603 \| 3.19 | 0.466 \| 3.15 | 0.711 \| 4.25 |
| | Homosc. + Corr. | 0.707 \| 4.96 | 0.953 \| 9.08 | 0.910 \| 7.25 | 0.542 \| 3.08 | 0.470 \| 3.41 | 0.632 \| 4.01 |
| | Heterosc. + Corr. | 0.685 \| 4.91 | 0.946 \| 11.33 | 0.939 \| 8.33 | 0.570 \| 2.90 | 0.400 \| 3.43 | 0.708 \| 4.07 |
| const = 0.3 | Homosc. + Ind. | 0.622 \| 3.81 | 0.939 \| 8.04 | 0.920 \| 7.48 | 0.653 \| 3.94 | 0.712 \| 4.73 | 0.654 \| 4.03 |
| | Heterosc. + Ind. | 0.663 \| 4.00 | 0.936 \| 9.48 | 0.922 \| 7.27 | 0.692 \| 3.83 | 0.780 \| 5.05 | 0.714 \| 3.86 |
| | Homosc. + Corr. | 0.594 \| 3.41 | 0.928 \| 7.45 | 0.912 \| 7.09 | 0.641 \| 3.78 | 0.765 \| 5.68 | 0.658 \| 4.05 |
| | Heterosc. + Corr. | 0.657 \| 3.48 | 0.937 \| 8.74 | 0.923 \| 6.80 | 0.653 \| 3.56 | 0.806 \| 6.17 | 0.729 \| 4.12 |
| const = 0.5 | Homosc. + Ind. | 0.619 \| 3.63 | 0.930 \| 7.56 | 0.912 \| 7.14 | 0.680 \| 4.15 | 0.729 \| 4.81 | 0.637 \| 3.86 |
| | Heterosc. + Ind. | 0.668 \| 3.63 | 0.936 \| 8.77 | 0.910 \| 6.59 | 0.717 \| 3.95 | 0.792 \| 4.98 | 0.718 \| 3.84 |
| | Homosc. + Corr. | 0.577 \| 3.26 | 0.927 \| 7.33 | 0.912 \| 7.03 | 0.663 \| 3.95 | 0.774 \| 5.49 | 0.643 \| 3.91 |
| | Heterosc. + Corr. | 0.641 \| 3.27 | 0.929 \| 7.95 | 0.916 \| 6.51 | 0.693 \| 3.75 | 0.815 \| 5.86 | 0.722 \| 3.90 |
| const = 0.7 | Homosc. + Ind. | 0.595 \| 3.40 | 0.929 \| 7.41 | 0.919 \| 7.21 | 0.658 \| 3.97 | 0.686 \| 4.37 | 0.654 \| 4.02 |
| | Heterosc. + Ind. | 0.673 \| 3.33 | 0.929 \| 8.16 | 0.911 \| 6.53 | 0.694 \| 3.81 | 0.762 \| 4.50 | 0.726 \| 4.07 |
| | Homosc. + Corr. | 0.594 \| 3.39 | 0.925 \| 7.31 | 0.910 \| 7.01 | 0.639 \| 3.77 | 0.722 \| 4.74 | 0.646 \| 3.96 |
| | Heterosc. + Corr. | 0.640 \| 3.29 | 0.935 \| 8.43 | 0.919 \| 6.69 | 0.655 \| 3.56 | 0.801 \| 5.20 | 0.729 \| 4.08 |
| const = 0.9 | Homosc. + Ind. | 0.604 \| 3.50 | 0.938 \| 7.75 | 0.923 \| 7.41 | 0.569 \| 3.34 | 0.568 \| 3.34 | 0.662 \| 4.21 |
| | Heterosc. + Ind. | 0.701 \| 4.20 | 0.943 \| 12.27 | 0.933 \| 8.03 | 0.626 \| 3.28 | 0.653 \| 3.39 | 0.702 \| 3.92 |
| | Homosc. + Corr. | 0.609 \| 3.54 | 0.920 \| 7.26 | 0.914 \| 7.12 | 0.550 \| 3.14 | 0.559 \| 3.18 | 0.646 \| 4.10 |
| | Heterosc. + Corr. | 0.666 \| 3.55 | 0.930 \| 9.26 | 0.943 \| 8.09 | 0.582 \| 2.90 | 0.666 \| 3.28 | 0.723 \| 4.59 |

### C.5 Detailed Results for Target Coverage 0.7

We report the detailed numerical results for target coverage $0.7$. The three tables in this subsection correspond to $\rho = 0.0$, $\rho = -0.5$, and $\rho = -1.0$, respectively.

Table 8: Mean realized coverage and mean interval length for target coverage 0.7 across propensity settings when $\rho = 0.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.685 \| 3.00 | 0.968 \| 6.43 | 0.961 \| 6.17 | 0.717 \| 3.24 | 0.772 \| 3.95 | 0.698 \| 3.27 |
| | Heterosc. + Ind. | 0.707 \| 2.90 | 0.941 \| 6.45 | 0.949 \| 5.64 | 0.726 \| 3.05 | 0.812 \| 4.08 | 0.699 \| 2.76 |
| | Homosc. + Corr. | 0.666 \| 2.82 | 0.965 \| 6.22 | 0.955 \| 5.92 | 0.706 \| 3.15 | 0.807 \| 4.49 | 0.698 \| 3.27 |
| | Heterosc. + Corr. | 0.718 \| 2.77 | 0.948 \| 6.40 | 0.952 \| 5.56 | 0.716 \| 2.95 | 0.836 \| 4.90 | 0.701 \| 2.68 |
| const = 0.1 | Homosc. + Ind. | 0.712 \| 3.88 | 0.975 \| 7.90 | 0.967 \| 7.22 | 0.587 \| 2.53 | 0.463 \| 2.40 | 0.674 \| 3.38 |
| | Heterosc. + Ind. | 0.734 \| 4.35 | 0.951 \| 8.79 | 0.951 \| 6.81 | 0.594 \| 2.45 | 0.396 \| 2.18 | 0.687 \| 2.94 |
| | Homosc. + Corr. | 0.734 \| 3.79 | 0.974 \| 7.34 | 0.968 \| 6.68 | 0.591 \| 2.50 | 0.411 \| 2.57 | 0.674 \| 3.38 |
| | Heterosc. + Corr. | 0.709 \| 3.72 | 0.951 \| 8.50 | 0.958 \| 6.64 | 0.595 \| 2.33 | 0.329 \| 2.40 | 0.689 \| 2.94 |
| const = 0.3 | Homosc. + Ind. | 0.681 \| 3.12 | 0.971 \| 6.81 | 0.961 \| 6.30 | 0.695 \| 3.11 | 0.729 \| 3.74 | 0.693 \| 3.27 |
| | Heterosc. + Ind. | 0.687 \| 3.25 | 0.941 \| 6.99 | 0.945 \| 5.71 | 0.697 \| 2.93 | 0.779 \| 3.94 | 0.698 \| 2.71 |
| | Homosc. + Corr. | 0.655 \| 2.80 | 0.972 \| 6.58 | 0.963 \| 6.23 | 0.691 \| 3.05 | 0.766 \| 4.42 | 0.702 \| 3.32 |
| | Heterosc. + Corr. | 0.730 \| 2.93 | 0.948 \| 6.45 | 0.949 \| 5.58 | 0.686 \| 2.88 | 0.802 \| 4.93 | 0.700 \| 2.73 |
| const = 0.5 | Homosc. + Ind. | 0.673 \| 2.91 | 0.962 \| 6.15 | 0.954 \| 5.93 | 0.728 \| 3.32 | 0.769 \| 3.89 | 0.691 \| 3.21 |
| | Heterosc. + Ind. | 0.690 \| 2.85 | 0.943 \| 6.53 | 0.945 \| 5.46 | 0.726 \| 3.09 | 0.820 \| 4.10 | 0.693 \| 2.64 |
| | Homosc. + Corr. | 0.656 \| 2.76 | 0.965 \| 6.20 | 0.961 \| 6.12 | 0.721 \| 3.23 | 0.816 \| 4.55 | 0.694 \| 3.23 |
| | Heterosc. + Corr. | 0.696 \| 2.71 | 0.948 \| 6.37 | 0.948 \| 5.45 | 0.724 \| 3.03 | 0.847 \| 5.01 | 0.694 \| 2.60 |
| const = 0.7 | Homosc. + Ind. | 0.668 \| 2.82 | 0.966 \| 6.25 | 0.959 \| 6.00 | 0.692 \| 3.09 | 0.731 \| 3.48 | 0.696 \| 3.30 |
| | Heterosc. + Ind. | 0.700 \| 2.57 | 0.946 \| 6.41 | 0.949 \| 5.50 | 0.704 \| 2.91 | 0.790 \| 3.64 | 0.676 \| 2.50 |
| | Homosc. + Corr. | 0.663 \| 2.81 | 0.968 \| 6.30 | 0.960 \| 6.07 | 0.691 \| 3.06 | 0.769 \| 3.86 | 0.705 \| 3.34 |
| | Heterosc. + Corr. | 0.697 \| 2.60 | 0.946 \| 6.33 | 0.942 \| 5.21 | 0.700 \| 2.91 | 0.821 \| 4.27 | 0.682 \| 2.55 |
| const = 0.9 | Homosc. + Ind. | 0.677 \| 2.91 | 0.975 \| 6.93 | 0.962 \| 6.18 | 0.616 \| 2.66 | 0.604 \| 2.54 | 0.678 \| 3.32 |
| | Heterosc. + Ind. | 0.702 \| 2.72 | 0.949 \| 7.71 | 0.952 \| 6.05 | 0.633 \| 2.49 | 0.661 \| 2.50 | 0.678 \| 2.74 |
| | Homosc. + Corr. | 0.654 \| 2.81 | 0.969 \| 6.54 | 0.964 \| 6.26 | 0.596 \| 2.54 | 0.615 \| 2.58 | 0.668 \| 3.30 |
| | Heterosc. + Corr. | 0.676 \| 2.44 | 0.936 \| 6.39 | 0.949 \| 5.79 | 0.601 \| 2.35 | 0.671 \| 2.47 | 0.694 \| 3.02 |

Table 9: Mean realized coverage and mean interval length for target coverage 0.7 across propensity settings when $\rho = -0.5$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.589 \| 2.97 | 0.923 \| 6.37 | 0.912 \| 6.13 | 0.630 \| 3.26 | 0.683 \| 3.88 | 0.613 \| 3.25 |
| | Heterosc. + Ind. | 0.653 \| 2.95 | 0.916 \| 6.92 | 0.905 \| 5.58 | 0.646 \| 3.01 | 0.755 \| 4.06 | 0.646 \| 2.65 |
| | Homosc. + Corr. | 0.572 \| 2.81 | 0.920 \| 6.20 | 0.908 \| 6.03 | 0.619 \| 3.14 | 0.736 \| 4.47 | 0.607 \| 3.19 |
| | Heterosc. + Corr. | 0.592 \| 2.61 | 0.914 \| 6.39 | 0.906 \| 5.47 | 0.632 \| 2.93 | 0.789 \| 4.94 | 0.653 \| 2.63 |
| const = 0.1 | Homosc. + Ind. | 0.668 \| 4.04 | 0.949 \| 8.00 | 0.924 \| 7.06 | 0.515 \| 2.56 | 0.412 \| 2.41 | 0.613 \| 3.43 |
| | Heterosc. + Ind. | 0.671 \| 4.15 | 0.912 \| 7.93 | 0.915 \| 7.17 | 0.533 \| 2.42 | 0.350 \| 2.16 | 0.642 \| 2.82 |
| | Homosc. + Corr. | 0.643 \| 3.66 | 0.942 \| 7.54 | 0.920 \| 6.46 | 0.513 \| 2.52 | 0.375 \| 2.57 | 0.603 \| 3.37 |
| | Heterosc. + Corr. | 0.652 \| 3.80 | 0.916 \| 7.92 | 0.916 \| 6.16 | 0.529 \| 2.32 | 0.296 \| 2.43 | 0.637 \| 2.78 |
| const = 0.3 | Homosc. + Ind. | 0.599 \| 3.15 | 0.931 \| 6.85 | 0.911 \| 6.21 | 0.604 \| 3.08 | 0.654 \| 3.75 | 0.608 \| 3.22 |
| | Heterosc. + Ind. | 0.622 \| 3.25 | 0.907 \| 6.86 | 0.905 \| 5.77 | 0.632 \| 2.94 | 0.724 \| 3.93 | 0.657 \| 2.77 |
| | Homosc. + Corr. | 0.592 \| 2.97 | 0.925 \| 6.38 | 0.917 \| 6.21 | 0.605 \| 3.06 | 0.702 \| 4.46 | 0.612 \| 3.27 |
| | Heterosc. + Corr. | 0.621 \| 2.87 | 0.920 \| 6.76 | 0.911 \| 5.62 | 0.608 \| 2.90 | 0.750 \| 4.91 | 0.656 \| 2.70 |
| const = 0.5 | Homosc. + Ind. | 0.579 \| 2.87 | 0.921 \| 6.33 | 0.911 \| 6.11 | 0.635 \| 3.28 | 0.691 \| 3.92 | 0.608 \| 3.20 |
| | Heterosc. + Ind. | 0.627 \| 2.83 | 0.903 \| 6.24 | 0.902 \| 5.41 | 0.656 \| 3.08 | 0.757 \| 4.07 | 0.641 \| 2.55 |
| | Homosc. + Corr. | 0.560 \| 2.73 | 0.924 \| 6.28 | 0.908 \| 6.01 | 0.629 \| 3.22 | 0.742 \| 4.52 | 0.619 \| 3.29 |
| | Heterosc. + Corr. | 0.608 \| 2.69 | 0.912 \| 6.14 | 0.906 \| 5.45 | 0.640 \| 3.04 | 0.794 \| 4.98 | 0.649 \| 2.58 |
| const = 0.7 | Homosc. + Ind. | 0.585 \| 2.87 | 0.918 \| 6.17 | 0.910 \| 6.05 | 0.611 \| 3.12 | 0.656 \| 3.54 | 0.603 \| 3.19 |
| | Heterosc. + Ind. | 0.641 \| 2.66 | 0.914 \| 6.33 | 0.909 \| 5.49 | 0.634 \| 2.92 | 0.733 \| 3.67 | 0.640 \| 2.55 |
| | Homosc. + Corr. | 0.545 \| 2.64 | 0.920 \| 6.20 | 0.914 \| 6.12 | 0.602 \| 3.05 | 0.686 \| 3.84 | 0.600 \| 3.17 |
| | Heterosc. + Corr. | 0.592 \| 2.56 | 0.913 \| 6.22 | 0.903 \| 5.35 | 0.616 \| 2.88 | 0.765 \| 4.24 | 0.641 \| 2.52 |
| const = 0.9 | Homosc. + Ind. | 0.564 \| 2.80 | 0.934 \| 6.69 | 0.921 \| 6.38 | 0.527 \| 2.62 | 0.520 \| 2.55 | 0.597 \| 3.26 |
| | Heterosc. + Ind. | 0.685 \| 3.23 | 0.917 \| 7.86 | 0.915 \| 5.89 | 0.560 \| 2.48 | 0.594 \| 2.47 | 0.626 \| 2.64 |
| | Homosc. + Corr. | 0.599 \| 2.98 | 0.941 \| 6.94 | 0.915 \| 6.19 | 0.512 \| 2.52 | 0.532 \| 2.59 | 0.615 \| 3.44 |
| | Heterosc. + Corr. | 0.624 \| 2.69 | 0.926 \| 7.90 | 0.907 \| 5.70 | 0.535 \| 2.34 | 0.607 \| 2.44 | 0.649 \| 2.86 |

Table 10: Mean realized coverage and mean interval length for target coverage 0.7 across propensity settings when $\rho = -1.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.522 \| 2.93 | 0.884 \| 6.53 | 0.872 \| 6.27 | 0.551 \| 3.14 | 0.616 \| 3.80 | 0.535 \| 3.12 |
| | Heterosc. + Ind. | 0.594 \| 2.86 | 0.886 \| 6.64 | 0.864 \| 5.44 | 0.582 \| 2.98 | 0.707 \| 4.02 | 0.601 \| 2.53 |
| | Homosc. + Corr. | 0.506 \| 2.77 | 0.875 \| 6.24 | 0.856 \| 5.98 | 0.554 \| 3.14 | 0.684 \| 4.51 | 0.554 \| 3.24 |
| | Heterosc. + Corr. | 0.559 \| 2.66 | 0.894 \| 6.75 | 0.859 \| 5.25 | 0.570 \| 3.00 | 0.754 \| 4.97 | 0.622 \| 2.66 |
| const = 0.1 | Homosc. + Ind. | 0.583 \| 3.77 | 0.914 \| 7.90 | 0.891 \| 7.22 | 0.460 \| 2.58 | 0.393 \| 2.52 | 0.557 \| 3.42 |
| | Heterosc. + Ind. | 0.572 \| 3.60 | 0.907 \| 9.03 | 0.885 \| 7.24 | 0.491 \| 2.43 | 0.338 \| 2.30 | 0.620 \| 3.05 |
| | Homosc. + Corr. | 0.583 \| 3.59 | 0.910 \| 7.65 | 0.864 \| 6.29 | 0.449 \| 2.47 | 0.354 \| 2.60 | 0.533 \| 3.21 |
| | Heterosc. + Corr. | 0.616 \| 3.77 | 0.909 \| 9.07 | 0.892 \| 6.43 | 0.464 \| 2.32 | 0.269 \| 2.44 | 0.608 \| 2.75 |
| const = 0.3 | Homosc. + Ind. | 0.542 \| 3.16 | 0.888 \| 6.74 | 0.874 \| 6.38 | 0.539 \| 3.07 | 0.597 \| 3.74 | 0.542 \| 3.17 |
| | Heterosc. + Ind. | 0.594 \| 3.26 | 0.886 \| 7.02 | 0.867 \| 5.69 | 0.589 \| 3.02 | 0.688 \| 3.96 | 0.612 \| 2.64 |
| | Homosc. + Corr. | 0.516 \| 2.87 | 0.879 \| 6.40 | 0.856 \| 6.01 | 0.538 \| 3.03 | 0.662 \| 4.52 | 0.552 \| 3.24 |
| | Heterosc. + Corr. | 0.566 \| 2.86 | 0.886 \| 6.46 | 0.876 \| 5.67 | 0.556 \| 2.90 | 0.730 \| 5.07 | 0.618 \| 2.72 |
| const = 0.5 | Homosc. + Ind. | 0.526 \| 2.94 | 0.879 \| 6.38 | 0.865 \| 6.12 | 0.575 \| 3.31 | 0.635 \| 3.96 | 0.542 \| 3.15 |
| | Heterosc. + Ind. | 0.568 \| 2.76 | 0.886 \| 6.54 | 0.861 \| 5.37 | 0.612 \| 3.12 | 0.717 \| 4.11 | 0.617 \| 2.63 |
| | Homosc. + Corr. | 0.508 \| 2.79 | 0.881 \| 6.35 | 0.862 \| 6.07 | 0.563 \| 3.19 | 0.692 \| 4.58 | 0.546 \| 3.18 |
| | Heterosc. + Corr. | 0.559 \| 2.71 | 0.883 \| 6.22 | 0.871 \| 5.52 | 0.582 \| 3.06 | 0.762 \| 5.09 | 0.613 \| 2.52 |
| const = 0.7 | Homosc. + Ind. | 0.503 \| 2.76 | 0.878 \| 6.32 | 0.852 \| 5.93 | 0.549 \| 3.13 | 0.593 \| 3.53 | 0.551 \| 3.24 |
| | Heterosc. + Ind. | 0.565 \| 2.58 | 0.880 \| 6.19 | 0.857 \| 5.25 | 0.584 \| 2.95 | 0.691 \| 3.72 | 0.616 \| 2.65 |
| | Homosc. + Corr. | 0.492 \| 2.69 | 0.878 \| 6.28 | 0.862 \| 6.07 | 0.534 \| 3.00 | 0.614 \| 3.69 | 0.538 \| 3.13 |
| | Heterosc. + Corr. | 0.552 \| 2.62 | 0.893 \| 6.50 | 0.873 \| 5.52 | 0.552 \| 2.87 | 0.730 \| 4.18 | 0.614 \| 2.62 |
| const = 0.9 | Homosc. + Ind. | 0.506 \| 2.81 | 0.890 \| 6.62 | 0.859 \| 6.11 | 0.465 \| 2.60 | 0.462 \| 2.57 | 0.563 \| 3.40 |
| | Heterosc. + Ind. | 0.587 \| 2.59 | 0.888 \| 7.30 | 0.882 \| 5.89 | 0.513 \| 2.47 | 0.544 \| 2.54 | 0.595 \| 2.57 |
| | Homosc. + Corr. | 0.509 \| 2.83 | 0.882 \| 6.45 | 0.871 \| 6.23 | 0.450 \| 2.47 | 0.462 \| 2.53 | 0.546 \| 3.32 |
| | Heterosc. + Corr. | 0.594 \| 2.85 | 0.894 \| 7.26 | 0.868 \| 5.66 | 0.470 \| 2.30 | 0.559 \| 2.42 | 0.612 \| 2.95 |

### C.6 Detailed Results for Target Coverage 0.6

We report the detailed numerical results for target coverage $0.6$. The three tables in this subsection correspond to $\rho = 0.0$, $\rho = -0.5$, and $\rho = -1.0$, respectively.

Table 11: Mean realized coverage and mean interval length for target coverage 0.6 across propensity settings when $\rho = 0.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.585 \| 2.40 | 0.944 \| 5.69 | 0.930 \| 5.35 | 0.615 \| 2.62 | 0.684 \| 3.30 | 0.593 \| 2.63 |
| | Heterosc. + Ind. | 0.617 \| 2.22 | 0.912 \| 5.34 | 0.916 \| 4.78 | 0.623 \| 2.44 | 0.735 \| 3.37 | 0.598 \| 1.90 |
| | Homosc. + Corr. | 0.564 \| 2.27 | 0.945 \| 5.62 | 0.942 \| 5.57 | 0.602 \| 2.53 | 0.719 \| 3.73 | 0.601 \| 2.66 |
| | Heterosc. + Corr. | 0.601 \| 2.22 | 0.909 \| 5.06 | 0.919 \| 4.85 | 0.607 \| 2.39 | 0.775 \| 4.20 | 0.597 \| 1.80 |
| const = 0.1 | Homosc. + Ind. | 0.624 \| 3.12 | 0.953 \| 7.13 | 0.934 \| 6.16 | 0.496 \| 2.04 | 0.378 \| 1.91 | 0.579 \| 2.72 |
| | Heterosc. + Ind. | 0.600 \| 3.18 | 0.925 \| 7.41 | 0.920 \| 6.00 | 0.494 \| 1.97 | 0.318 \| 1.71 | 0.587 \| 2.11 |
| | Homosc. + Corr. | 0.592 \| 2.68 | 0.945 \| 6.37 | 0.938 \| 5.71 | 0.498 \| 2.03 | 0.328 \| 2.04 | 0.591 \| 2.82 |
| | Heterosc. + Corr. | 0.681 \| 3.24 | 0.929 \| 6.93 | 0.915 \| 4.98 | 0.501 \| 1.87 | 0.257 \| 1.84 | 0.600 \| 2.12 |
| const = 0.3 | Homosc. + Ind. | 0.609 \| 2.65 | 0.944 \| 5.87 | 0.932 \| 5.49 | 0.584 \| 2.44 | 0.588 \| 2.91 | 0.592 \| 2.63 |
| | Heterosc. + Ind. | 0.629 \| 2.69 | 0.913 \| 5.75 | 0.912 \| 4.80 | 0.592 \| 2.34 | 0.630 \| 3.04 | 0.600 \| 1.89 |
| | Homosc. + Corr. | 0.602 \| 2.49 | 0.940 \| 5.57 | 0.928 \| 5.36 | 0.582 \| 2.41 | 0.642 \| 3.53 | 0.606 \| 2.72 |
| | Heterosc. + Corr. | 0.634 \| 2.35 | 0.911 \| 5.15 | 0.911 \| 4.63 | 0.579 \| 2.33 | 0.678 \| 3.87 | 0.593 \| 1.81 |
| const = 0.5 | Homosc. + Ind. | 0.574 \| 2.33 | 0.938 \| 5.48 | 0.931 \| 5.33 | 0.620 \| 2.62 | 0.673 \| 3.22 | 0.589 \| 2.59 |
| | Heterosc. + Ind. | 0.608 \| 2.18 | 0.903 \| 4.95 | 0.909 \| 4.56 | 0.617 \| 2.48 | 0.741 \| 3.39 | 0.587 \| 1.79 |
| | Homosc. + Corr. | 0.578 \| 2.34 | 0.945 \| 5.61 | 0.935 \| 5.42 | 0.617 \| 2.59 | 0.729 \| 3.76 | 0.595 \| 2.63 |
| | Heterosc. + Corr. | 0.608 \| 2.25 | 0.915 \| 5.19 | 0.917 \| 4.72 | 0.620 \| 2.46 | 0.784 \| 4.27 | 0.603 \| 1.82 |
| const = 0.7 | Homosc. + Ind. | 0.581 \| 2.34 | 0.946 \| 5.67 | 0.936 \| 5.40 | 0.588 \| 2.46 | 0.637 \| 2.83 | 0.600 \| 2.67 |
| | Heterosc. + Ind. | 0.617 \| 2.09 | 0.909 \| 5.00 | 0.910 \| 4.49 | 0.597 \| 2.33 | 0.708 \| 2.98 | 0.586 \| 1.81 |
| | Homosc. + Corr. | 0.583 \| 2.36 | 0.940 \| 5.50 | 0.929 \| 5.28 | 0.580 \| 2.41 | 0.669 \| 3.09 | 0.595 \| 2.66 |
| | Heterosc. + Corr. | 0.613 \| 2.16 | 0.909 \| 4.99 | 0.904 \| 4.43 | 0.588 \| 2.30 | 0.752 \| 3.50 | 0.576 \| 1.69 |
| const = 0.9 | Homosc. + Ind. | 0.615 \| 2.58 | 0.941 \| 5.69 | 0.942 \| 5.68 | 0.516 \| 2.12 | 0.508 \| 2.03 | 0.571 \| 2.65 |
| | Heterosc. + Ind. | 0.645 \| 2.17 | 0.909 \| 5.53 | 0.927 \| 5.34 | 0.537 \| 1.99 | 0.569 \| 1.97 | 0.581 \| 1.97 |
| | Homosc. + Corr. | 0.576 \| 2.35 | 0.955 \| 5.99 | 0.939 \| 5.55 | 0.499 \| 2.03 | 0.521 \| 2.08 | 0.579 \| 2.72 |
| | Heterosc. + Corr. | 0.645 \| 2.25 | 0.909 \| 5.79 | 0.917 \| 4.83 | 0.513 \| 1.91 | 0.559 \| 1.87 | 0.604 \| 2.13 |

Table 12: Mean realized coverage and mean interval length for target coverage 0.6 across propensity settings when $\rho = -0.5$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.501 \| 2.41 | 0.885 \| 5.65 | 0.872 \| 5.42 | 0.526 \| 2.59 | 0.590 \| 3.21 | 0.518 \| 2.63 |
| | Heterosc. + Ind. | 0.548 \| 2.22 | 0.876 \| 5.47 | 0.860 \| 4.61 | 0.548 \| 2.43 | 0.672 \| 3.36 | 0.543 \| 1.82 |
| | Homosc. + Corr. | 0.494 \| 2.35 | 0.886 \| 5.64 | 0.875 \| 5.47 | 0.511 \| 2.49 | 0.651 \| 3.76 | 0.518 \| 2.62 |
| | Heterosc. + Corr. | 0.543 \| 2.22 | 0.879 \| 5.39 | 0.872 \| 4.86 | 0.530 \| 2.37 | 0.726 \| 4.23 | 0.548 \| 1.75 |
| const = 0.1 | Homosc. + Ind. | 0.580 \| 3.32 | 0.925 \| 7.49 | 0.877 \| 6.01 | 0.420 \| 2.02 | 0.330 \| 1.90 | 0.520 \| 2.77 |
| | Heterosc. + Ind. | 0.559 \| 3.21 | 0.891 \| 7.67 | 0.878 \| 5.87 | 0.433 \| 1.92 | 0.278 \| 1.71 | 0.534 \| 1.96 |
| | Homosc. + Corr. | 0.547 \| 2.86 | 0.916 \| 6.72 | 0.884 \| 5.75 | 0.422 \| 2.02 | 0.296 \| 2.04 | 0.511 \| 2.75 |
| | Heterosc. + Corr. | 0.584 \| 3.12 | 0.900 \| 7.22 | 0.878 \| 5.12 | 0.428 \| 1.84 | 0.221 \| 1.84 | 0.555 \| 2.07 |
| const = 0.3 | Homosc. + Ind. | 0.519 \| 2.58 | 0.881 \| 5.73 | 0.870 \| 5.43 | 0.502 \| 2.45 | 0.522 \| 2.92 | 0.515 \| 2.62 |
| | Heterosc. + Ind. | 0.564 \| 2.63 | 0.870 \| 5.65 | 0.869 \| 4.86 | 0.532 \| 2.36 | 0.571 \| 3.04 | 0.540 \| 1.78 |
| | Homosc. + Corr. | 0.482 \| 2.30 | 0.877 \| 5.52 | 0.878 \| 5.51 | 0.502 \| 2.43 | 0.577 \| 3.54 | 0.520 \| 2.67 |
| | Heterosc. + Corr. | 0.521 \| 2.32 | 0.875 \| 5.36 | 0.863 \| 4.68 | 0.504 \| 2.36 | 0.635 \| 3.91 | 0.549 \| 1.79 |
| const = 0.5 | Homosc. + Ind. | 0.496 \| 2.37 | 0.878 \| 5.51 | 0.866 \| 5.32 | 0.538 \| 2.65 | 0.597 \| 3.25 | 0.515 \| 2.61 |
| | Heterosc. + Ind. | 0.543 \| 2.19 | 0.875 \| 5.38 | 0.863 \| 4.63 | 0.553 \| 2.48 | 0.680 \| 3.39 | 0.542 \| 1.80 |
| | Homosc. + Corr. | 0.484 \| 2.29 | 0.884 \| 5.57 | 0.871 \| 5.39 | 0.529 \| 2.58 | 0.646 \| 3.72 | 0.519 \| 2.65 |
| | Heterosc. + Corr. | 0.522 \| 2.25 | 0.871 \| 5.08 | 0.868 \| 4.72 | 0.536 \| 2.46 | 0.730 \| 4.27 | 0.551 \| 1.77 |
| const = 0.7 | Homosc. + Ind. | 0.501 \| 2.38 | 0.887 \| 5.61 | 0.869 \| 5.37 | 0.511 \| 2.50 | 0.555 \| 2.83 | 0.518 \| 2.64 |
| | Heterosc. + Ind. | 0.546 \| 2.10 | 0.867 \| 4.99 | 0.856 \| 4.49 | 0.531 \| 2.32 | 0.654 \| 3.01 | 0.548 \| 1.84 |
| | Homosc. + Corr. | 0.486 \| 2.30 | 0.878 \| 5.47 | 0.866 \| 5.29 | 0.500 \| 2.42 | 0.584 \| 3.07 | 0.512 \| 2.61 |
| | Heterosc. + Corr. | 0.512 \| 2.15 | 0.868 \| 5.03 | 0.862 \| 4.59 | 0.513 \| 2.32 | 0.690 \| 3.45 | 0.539 \| 1.72 |
| const = 0.9 | Homosc. + Ind. | 0.526 \| 2.55 | 0.901 \| 5.97 | 0.875 \| 5.47 | 0.436 \| 2.09 | 0.431 \| 2.04 | 0.515 \| 2.73 |
| | Heterosc. + Ind. | 0.577 \| 2.34 | 0.873 \| 5.95 | 0.874 \| 5.05 | 0.469 \| 1.99 | 0.498 \| 1.96 | 0.532 \| 1.92 |
| | Homosc. + Corr. | 0.486 \| 2.31 | 0.902 \| 6.00 | 0.874 \| 5.50 | 0.423 \| 2.03 | 0.433 \| 2.04 | 0.507 \| 2.69 |
| | Heterosc. + Corr. | 0.563 \| 2.23 | 0.880 \| 5.91 | 0.855 \| 4.58 | 0.437 \| 1.89 | 0.494 \| 1.88 | 0.564 \| 2.03 |

Table 13: Mean realized coverage and mean interval length for target coverage 0.6 across propensity settings when $\rho = -1.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.429 \| 2.32 | 0.827 \| 5.58 | 0.806 \| 5.32 | 0.459 \| 2.53 | 0.525 \| 3.13 | 0.442 \| 2.49 |
| | Heterosc. + Ind. | 0.512 \| 2.25 | 0.841 \| 5.43 | 0.804 \| 4.48 | 0.489 \| 2.42 | 0.628 \| 3.34 | 0.500 \| 1.75 |
| | Homosc. + Corr. | 0.434 \| 2.33 | 0.825 \| 5.52 | 0.805 \| 5.27 | 0.463 \| 2.54 | 0.593 \| 3.75 | 0.467 \| 2.65 |
| | Heterosc. + Corr. | 0.483 \| 2.21 | 0.833 \| 5.07 | 0.821 \| 4.67 | 0.479 \| 2.41 | 0.694 \| 4.24 | 0.517 \| 1.78 |
| const = 0.1 | Homosc. + Ind. | 0.517 \| 3.20 | 0.856 \| 6.79 | 0.835 \| 6.11 | 0.379 \| 2.06 | 0.311 \| 1.98 | 0.461 \| 2.73 |
| | Heterosc. + Ind. | 0.535 \| 3.26 | 0.865 \| 8.14 | 0.834 \| 5.77 | 0.403 \| 1.94 | 0.267 \| 1.81 | 0.513 \| 2.13 |
| | Homosc. + Corr. | 0.460 \| 2.67 | 0.849 \| 6.30 | 0.818 \| 5.59 | 0.369 \| 1.99 | 0.273 \| 2.02 | 0.467 \| 2.75 |
| | Heterosc. + Corr. | 0.543 \| 3.01 | 0.865 \| 6.86 | 0.830 \| 5.09 | 0.379 \| 1.88 | 0.210 \| 1.91 | 0.512 \| 1.94 |
| const = 0.3 | Homosc. + Ind. | 0.470 \| 2.62 | 0.845 \| 5.93 | 0.812 \| 5.42 | 0.446 \| 2.46 | 0.470 \| 2.89 | 0.456 \| 2.59 |
| | Heterosc. + Ind. | 0.518 \| 2.61 | 0.835 \| 5.61 | 0.819 \| 4.78 | 0.483 \| 2.40 | 0.536 \| 3.07 | 0.520 \| 1.88 |
| | Homosc. + Corr. | 0.423 \| 2.27 | 0.830 \| 5.62 | 0.815 \| 5.42 | 0.446 \| 2.43 | 0.529 \| 3.51 | 0.461 \| 2.62 |
| | Heterosc. + Corr. | 0.498 \| 2.36 | 0.835 \| 5.11 | 0.825 \| 4.78 | 0.461 \| 2.36 | 0.587 \| 3.90 | 0.510 \| 1.82 |
| const = 0.5 | Homosc. + Ind. | 0.436 \| 2.36 | 0.820 \| 5.46 | 0.811 \| 5.37 | 0.480 \| 2.66 | 0.542 \| 3.26 | 0.458 \| 2.58 |
| | Heterosc. + Ind. | 0.501 \| 2.23 | 0.829 \| 5.04 | 0.810 \| 4.52 | 0.512 \| 2.52 | 0.644 \| 3.44 | 0.514 \| 1.80 |
| | Homosc. + Corr. | 0.428 \| 2.29 | 0.821 \| 5.45 | 0.809 \| 5.32 | 0.469 \| 2.57 | 0.598 \| 3.76 | 0.458 \| 2.58 |
| | Heterosc. + Corr. | 0.480 \| 2.26 | 0.833 \| 5.04 | 0.822 \| 4.68 | 0.487 \| 2.48 | 0.699 \| 4.32 | 0.509 \| 1.69 |
| const = 0.7 | Homosc. + Ind. | 0.425 \| 2.28 | 0.826 \| 5.52 | 0.809 \| 5.33 | 0.455 \| 2.50 | 0.501 \| 2.88 | 0.452 \| 2.57 |
| | Heterosc. + Ind. | 0.469 \| 2.05 | 0.828 \| 4.95 | 0.816 \| 4.58 | 0.488 \| 2.36 | 0.605 \| 3.02 | 0.505 \| 1.76 |
| | Homosc. + Corr. | 0.434 \| 2.33 | 0.826 \| 5.54 | 0.807 \| 5.31 | 0.444 \| 2.42 | 0.513 \| 2.93 | 0.452 \| 2.56 |
| | Heterosc. + Corr. | 0.468 \| 2.14 | 0.829 \| 4.92 | 0.819 \| 4.63 | 0.452 \| 2.33 | 0.654 \| 3.35 | 0.511 \| 1.73 |
| const = 0.9 | Homosc. + Ind. | 0.449 \| 2.44 | 0.848 \| 5.95 | 0.807 \| 5.34 | 0.381 \| 2.08 | 0.379 \| 2.05 | 0.476 \| 2.78 |
| | Heterosc. + Ind. | 0.540 \| 2.21 | 0.837 \| 5.54 | 0.829 \| 4.95 | 0.425 \| 1.97 | 0.458 \| 1.99 | 0.501 \| 1.83 |
| | Homosc. + Corr. | 0.461 \| 2.51 | 0.833 \| 5.68 | 0.817 \| 5.44 | 0.369 \| 1.98 | 0.379 \| 2.01 | 0.453 \| 2.65 |
| | Heterosc. + Corr. | 0.504 \| 2.20 | 0.824 \| 5.11 | 0.810 \| 4.70 | 0.384 \| 1.84 | 0.453 \| 1.86 | 0.529 \| 2.10 |

### C.7 Detailed Results for Target Coverage 0.5

We report the detailed numerical results for target coverage $0.5$. The three tables in this subsection correspond to $\rho = 0.0$, $\rho = -0.5$, and $\rho = -1.0$, respectively.

Table 14: Mean realized coverage and mean interval length for target coverage 0.5 across propensity settings when $\rho = 0.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.530 \| 2.12 | 0.913 \| 5.04 | 0.898 \| 4.79 | 0.517 \| 2.11 | 0.578 \| 2.67 | 0.494 \| 2.11 |
| | Heterosc. + Ind. | 0.527 \| 1.76 | 0.878 \| 4.54 | 0.868 \| 3.93 | 0.520 \| 1.95 | 0.650 \| 2.80 | 0.500 \| 1.34 |
| | Homosc. + Corr. | 0.504 \| 1.98 | 0.917 \| 5.07 | 0.904 \| 4.88 | 0.503 \| 2.02 | 0.627 \| 3.09 | 0.510 \| 2.18 |
| | Heterosc. + Corr. | 0.539 \| 1.87 | 0.878 \| 4.28 | 0.863 \| 3.81 | 0.502 \| 1.91 | 0.705 \| 3.61 | 0.507 \| 1.27 |
| const = 0.1 | Homosc. + Ind. | 0.553 \| 2.75 | 0.908 \| 5.97 | 0.896 \| 5.27 | 0.406 \| 1.62 | 0.296 \| 1.47 | 0.472 \| 2.13 |
| | Heterosc. + Ind. | 0.571 \| 2.94 | 0.887 \| 6.17 | 0.878 \| 4.88 | 0.393 \| 1.52 | 0.249 \| 1.34 | 0.467 \| 1.42 |
| | Homosc. + Corr. | 0.483 \| 2.06 | 0.909 \| 5.40 | 0.894 \| 4.90 | 0.409 \| 1.62 | 0.262 \| 1.63 | 0.491 \| 2.23 |
| | Heterosc. + Corr. | 0.565 \| 2.32 | 0.885 \| 5.40 | 0.870 \| 4.17 | 0.408 \| 1.52 | 0.199 \| 1.46 | 0.494 \| 1.44 |
| const = 0.3 | Homosc. + Ind. | 0.521 \| 2.12 | 0.915 \| 5.21 | 0.901 \| 4.87 | 0.479 \| 1.92 | 0.459 \| 2.24 | 0.501 \| 2.15 |
| | Heterosc. + Ind. | 0.558 \| 2.08 | 0.865 \| 4.47 | 0.878 \| 4.15 | 0.484 \| 1.86 | 0.452 \| 2.29 | 0.491 \| 1.31 |
| | Homosc. + Corr. | 0.497 \| 1.97 | 0.910 \| 5.03 | 0.900 \| 4.86 | 0.483 \| 1.92 | 0.496 \| 2.75 | 0.498 \| 2.15 |
| | Heterosc. + Corr. | 0.522 \| 1.95 | 0.868 \| 4.24 | 0.879 \| 4.10 | 0.478 \| 1.86 | 0.504 \| 3.04 | 0.501 \| 1.28 |
| const = 0.5 | Homosc. + Ind. | 0.516 \| 2.03 | 0.906 \| 4.90 | 0.897 \| 4.78 | 0.521 \| 2.10 | 0.575 \| 2.63 | 0.487 \| 2.06 |
| | Heterosc. + Ind. | 0.520 \| 1.76 | 0.867 \| 4.19 | 0.863 \| 3.81 | 0.523 \| 2.00 | 0.658 \| 2.86 | 0.491 \| 1.29 |
| | Homosc. + Corr. | 0.518 \| 2.04 | 0.912 \| 4.97 | 0.896 \| 4.78 | 0.511 \| 2.05 | 0.621 \| 3.03 | 0.492 \| 2.09 |
| | Heterosc. + Corr. | 0.528 \| 1.91 | 0.869 \| 4.15 | 0.875 \| 3.99 | 0.515 \| 1.97 | 0.716 \| 3.62 | 0.498 \| 1.22 |
| const = 0.7 | Homosc. + Ind. | 0.512 \| 2.01 | 0.902 \| 4.79 | 0.904 \| 4.84 | 0.491 \| 1.97 | 0.534 \| 2.24 | 0.497 \| 2.13 |
| | Heterosc. + Ind. | 0.548 \| 1.74 | 0.868 \| 4.28 | 0.870 \| 3.90 | 0.507 \| 1.88 | 0.620 \| 2.42 | 0.499 \| 1.32 |
| | Homosc. + Corr. | 0.514 \| 2.02 | 0.897 \| 4.76 | 0.893 \| 4.71 | 0.474 \| 1.89 | 0.566 \| 2.43 | 0.489 \| 2.09 |
| | Heterosc. + Corr. | 0.537 \| 1.81 | 0.858 \| 3.91 | 0.872 \| 3.99 | 0.491 \| 1.85 | 0.668 \| 2.81 | 0.482 \| 1.16 |
| const = 0.9 | Homosc. + Ind. | 0.514 \| 2.06 | 0.917 \| 5.26 | 0.896 \| 4.77 | 0.430 \| 1.71 | 0.423 \| 1.63 | 0.464 \| 2.07 |
| | Heterosc. + Ind. | 0.551 \| 1.68 | 0.886 \| 5.03 | 0.880 \| 4.16 | 0.445 \| 1.57 | 0.478 \| 1.56 | 0.500 \| 1.49 |
| | Homosc. + Corr. | 0.558 \| 2.28 | 0.900 \| 4.89 | 0.907 \| 4.95 | 0.411 \| 1.63 | 0.422 \| 1.62 | 0.471 \| 2.13 |
| | Heterosc. + Corr. | 0.589 \| 1.90 | 0.862 \| 4.43 | 0.889 \| 4.49 | 0.419 \| 1.52 | 0.457 \| 1.49 | 0.485 \| 1.34 |

Table 15: Mean realized coverage and mean interval length for target coverage 0.5 across propensity settings when $\rho = -0.5$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.439 \| 2.07 | 0.834 \| 4.93 | 0.822 \| 4.78 | 0.433 \| 2.07 | 0.501 \| 2.65 | 0.426 \| 2.10 |
| | Heterosc. + Ind. | 0.475 \| 1.80 | 0.823 \| 4.36 | 0.809 \| 3.91 | 0.453 \| 1.94 | 0.596 \| 2.81 | 0.449 \| 1.30 |
| | Homosc. + Corr. | 0.414 \| 1.92 | 0.837 \| 4.92 | 0.821 \| 4.77 | 0.417 \| 1.96 | 0.560 \| 3.11 | 0.421 \| 2.07 |
| | Heterosc. + Corr. | 0.445 \| 1.83 | 0.824 \| 4.27 | 0.798 \| 3.79 | 0.432 \| 1.93 | 0.656 \| 3.63 | 0.453 \| 1.21 |
| const = 0.1 | Homosc. + Ind. | 0.481 \| 2.63 | 0.881 \| 6.46 | 0.838 \| 5.34 | 0.346 \| 1.63 | 0.264 \| 1.50 | 0.421 \| 2.17 |
| | Heterosc. + Ind. | 0.499 \| 2.74 | 0.861 \| 6.31 | 0.838 \| 5.04 | 0.345 \| 1.49 | 0.220 \| 1.35 | 0.443 \| 1.43 |
| | Homosc. + Corr. | 0.467 \| 2.34 | 0.873 \| 5.78 | 0.822 \| 4.91 | 0.342 \| 1.60 | 0.232 \| 1.61 | 0.404 \| 2.09 |
| | Heterosc. + Corr. | 0.518 \| 2.41 | 0.849 \| 5.25 | 0.844 \| 4.73 | 0.343 \| 1.47 | 0.172 \| 1.47 | 0.459 \| 1.47 |
| const = 0.3 | Homosc. + Ind. | 0.439 \| 2.11 | 0.836 \| 5.05 | 0.824 \| 4.83 | 0.412 \| 1.95 | 0.405 \| 2.25 | 0.423 \| 2.08 |
| | Heterosc. + Ind. | 0.462 \| 1.99 | 0.817 \| 4.46 | 0.815 \| 4.04 | 0.424 \| 1.84 | 0.408 \| 2.31 | 0.445 \| 1.26 |
| | Homosc. + Corr. | 0.425 \| 1.99 | 0.835 \| 4.96 | 0.838 \| 4.98 | 0.412 \| 1.94 | 0.449 \| 2.76 | 0.439 \| 2.19 |
| | Heterosc. + Corr. | 0.464 \| 1.94 | 0.821 \| 4.23 | 0.820 \| 4.12 | 0.408 \| 1.88 | 0.459 \| 3.03 | 0.454 \| 1.24 |
| const = 0.5 | Homosc. + Ind. | 0.420 \| 1.95 | 0.834 \| 4.90 | 0.812 \| 4.65 | 0.438 \| 2.08 | 0.497 \| 2.61 | 0.426 \| 2.11 |
| | Heterosc. + Ind. | 0.451 \| 1.77 | 0.811 \| 4.08 | 0.814 \| 3.94 | 0.464 \| 2.00 | 0.597 \| 2.85 | 0.446 \| 1.26 |
| | Homosc. + Corr. | 0.425 \| 1.97 | 0.842 \| 4.98 | 0.821 \| 4.77 | 0.438 \| 2.07 | 0.551 \| 3.04 | 0.425 \| 2.11 |
| | Heterosc. + Corr. | 0.467 \| 1.91 | 0.825 \| 4.27 | 0.817 \| 4.01 | 0.442 \| 1.98 | 0.661 \| 3.61 | 0.455 \| 1.22 |
| const = 0.7 | Homosc. + Ind. | 0.425 \| 1.97 | 0.833 \| 4.88 | 0.818 \| 4.71 | 0.417 \| 1.98 | 0.466 \| 2.29 | 0.429 \| 2.13 |
| | Heterosc. + Ind. | 0.467 \| 1.71 | 0.823 \| 4.25 | 0.821 \| 4.03 | 0.439 \| 1.85 | 0.570 \| 2.44 | 0.456 \| 1.31 |
| | Homosc. + Corr. | 0.418 \| 1.94 | 0.834 \| 4.88 | 0.825 \| 4.81 | 0.404 \| 1.90 | 0.484 \| 2.41 | 0.424 \| 2.10 |
| | Heterosc. + Corr. | 0.484 \| 1.86 | 0.822 \| 4.17 | 0.809 \| 3.91 | 0.417 \| 1.84 | 0.618 \| 2.81 | 0.448 \| 1.19 |
| const = 0.9 | Homosc. + Ind. | 0.438 \| 2.05 | 0.851 \| 5.21 | 0.827 \| 4.87 | 0.356 \| 1.67 | 0.352 \| 1.62 | 0.434 \| 2.24 |
| | Heterosc. + Ind. | 0.506 \| 1.81 | 0.846 \| 5.03 | 0.804 \| 3.87 | 0.383 \| 1.55 | 0.417 \| 1.56 | 0.450 \| 1.42 |
| | Homosc. + Corr. | 0.444 \| 2.08 | 0.849 \| 5.14 | 0.821 \| 4.76 | 0.347 \| 1.63 | 0.352 \| 1.62 | 0.419 \| 2.16 |
| | Heterosc. + Corr. | 0.474 \| 1.72 | 0.829 \| 4.70 | 0.809 \| 3.92 | 0.357 \| 1.53 | 0.402 \| 1.49 | 0.462 \| 1.41 |

Table 16: Mean realized coverage and mean interval length for target coverage 0.5 across propensity settings when $\rho = -1.0$. Entries are coverage | length.

| Propensity | Scenario | Nest-Inexact | Nest-Exact | Naive | DR | IPW | Split-Pair |
|---|---|---|---|---|---|---|---|
| beta | Homosc. + Ind. | 0.370 \| 1.98 | 0.773 \| 4.93 | 0.763 \| 4.83 | 0.377 \| 2.03 | 0.439 \| 2.56 | 0.354 \| 1.96 |
| | Heterosc. + Ind. | 0.420 \| 1.76 | 0.780 \| 4.24 | 0.769 \| 3.95 | 0.395 \| 1.92 | 0.548 \| 2.79 | 0.410 \| 1.25 |
| | Homosc. + Corr. | 0.367 \| 1.93 | 0.776 \| 4.94 | 0.762 \| 4.81 | 0.378 \| 2.02 | 0.503 \| 3.07 | 0.383 \| 2.12 |
| | Heterosc. + Corr. | 0.410 \| 1.85 | 0.773 \| 4.03 | 0.776 \| 4.08 | 0.386 \| 1.93 | 0.619 \| 3.60 | 0.427 \| 1.24 |
| const = 0.1 | Homosc. + Ind. | 0.468 \| 2.84 | 0.781 \| 5.60 | 0.792 \| 5.45 | 0.310 \| 1.66 | 0.249 \| 1.57 | 0.366 \| 2.11 |
| | Heterosc. + Ind. | 0.497 \| 2.93 | 0.819 \| 6.03 | 0.792 \| 4.82 | 0.324 \| 1.52 | 0.207 \| 1.39 | 0.418 \| 1.53 |
| | Homosc. + Corr. | 0.431 \| 2.43 | 0.803 \| 5.55 | 0.773 \| 5.01 | 0.296 \| 1.57 | 0.211 \| 1.58 | 0.375 \| 2.17 |
| | Heterosc. + Corr. | 0.454 \| 2.29 | 0.810 \| 5.36 | 0.786 \| 4.46 | 0.303 \| 1.50 | 0.157 \| 1.49 | 0.432 \| 1.45 |
| const = 0.3 | Homosc. + Ind. | 0.391 \| 2.13 | 0.776 \| 5.04 | 0.759 \| 4.78 | 0.363 \| 1.95 | 0.367 \| 2.25 | 0.374 \| 2.07 |
| | Heterosc. + Ind. | 0.421 \| 1.92 | 0.777 \| 4.41 | 0.772 \| 4.03 | 0.387 \| 1.90 | 0.372 \| 2.33 | 0.409 \| 1.25 |
| | Homosc. + Corr. | 0.379 \| 2.01 | 0.785 \| 5.07 | 0.761 \| 4.80 | 0.366 \| 1.95 | 0.410 \| 2.72 | 0.386 \| 2.15 |
| | Heterosc. + Corr. | 0.414 \| 1.93 | 0.789 \| 4.32 | 0.766 \| 3.94 | 0.378 \| 1.89 | 0.401 \| 2.95 | 0.409 \| 1.19 |
| const = 0.5 | Homosc. + Ind. | 0.375 \| 1.99 | 0.773 \| 4.92 | 0.758 \| 4.75 | 0.396 \| 2.15 | 0.454 \| 2.66 | 0.371 \| 2.05 |
| | Heterosc. + Ind. | 0.417 \| 1.77 | 0.791 \| 4.37 | 0.761 \| 3.88 | 0.423 \| 2.01 | 0.571 \| 2.91 | 0.417 \| 1.24 |
| | Homosc. + Corr. | 0.380 \| 2.00 | 0.772 \| 4.90 | 0.768 \| 4.86 | 0.386 \| 2.06 | 0.519 \| 3.17 | 0.367 \| 2.02 |
| | Heterosc. + Corr. | 0.416 \| 1.88 | 0.769 \| 3.98 | 0.751 \| 3.84 | 0.395 \| 1.99 | 0.629 \| 3.69 | 0.411 \| 1.14 |
| const = 0.7 | Homosc. + Ind. | 0.377 \| 2.00 | 0.765 \| 4.81 | 0.760 \| 4.77 | 0.368 \| 1.98 | 0.414 \| 2.31 | 0.367 \| 2.04 |
| | Heterosc. + Ind. | 0.418 \| 1.71 | 0.774 \| 4.08 | 0.762 \| 3.86 | 0.399 \| 1.87 | 0.532 \| 2.49 | 0.415 \| 1.26 |
| | Homosc. + Corr. | 0.371 \| 1.95 | 0.762 \| 4.79 | 0.769 \| 4.87 | 0.363 \| 1.93 | 0.424 \| 2.33 | 0.369 \| 2.04 |
| | Heterosc. + Corr. | 0.404 \| 1.81 | 0.783 \| 4.19 | 0.758 \| 3.86 | 0.373 \| 1.87 | 0.579 \| 2.70 | 0.419 \| 1.20 |
| const = 0.9 | Homosc. + Ind. | 0.385 \| 2.05 | 0.784 \| 5.06 | 0.743 \| 4.62 | 0.306 \| 1.64 | 0.302 \| 1.59 | 0.382 \| 2.17 |
| | Heterosc. + Ind. | 0.415 \| 1.57 | 0.788 \| 4.56 | 0.781 \| 4.16 | 0.341 \| 1.54 | 0.371 \| 1.53 | 0.430 \| 1.40 |
| | Homosc. + Corr. | 0.383 \| 2.04 | 0.790 \| 5.14 | 0.785 \| 5.08 | 0.298 \| 1.58 | 0.309 \| 1.61 | 0.376 \| 2.16 |
| | Heterosc. + Corr. | 0.428 \| 1.75 | 0.775 \| 4.28 | 0.770 \| 4.06 | 0.311 \| 1.46 | 0.361 \| 1.45 | 0.430 \| 1.38 |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper. The paper clearly states its impossibility results for distribution-free predictive inference of ITEs under standard causal assumptions, its positive result under additional structural assumptions, and its empirical illustrations in Sections 2–4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the scope and limitations of the results. In particular, the infinite-expected-length conclusion for distribution-free prediction of realized ITEs requires no regularity assumption on the covariates and holds even in their absence, provided coverage holds uniformly over the corresponding strong-overlap class. The stronger pointwise inclusion-probability results use an absolutely continuous covariate coordinate. The positive Split-Pair result requires additional structural assumptions, including covariate-independent residuals and conditional independence between potential outcomes.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are stated with their assumptions in Section 2 and Appendix A. Complete proofs of the impossibility results, supporting lemmas, and the positive Split-Pair result are provided in Appendix A, with the main proof ingredients and external results properly referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main simulation results. The data-generating processes, propensity score regimes, error dependence structures, compared methods, evaluation metrics, and additional implementation details are described in Section 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides anonymized simulation code and reproduction instructions in the supplemental material. All experiments are based on synthetic data generated from the data-generating processes described in Section 4 and Appendix C, and the code will be de-anonymized and publicly released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings needed to understand the simulation results. The data-generating processes, treatment assignment mechanisms, error dependence structures, data splitting schemes, compared methods, evaluation metrics, and additional implementation details are described in Section 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports standard errors for the main simulation figures that support the empirical conclusions. The standard errors are computed across Monte Carlo replications under each experimental condition, as described in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper does not provide detailed compute-resource information for each individual experiment. The experiments are lightweight simulation studies used to illustrate the theoretical findings, and all reported experiments can be run on a standard laptop CPU without specialized hardware.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. The paper is theoretical and simulation-based, does not involve human subjects, private or sensitive data, or high-risk deployment, and the submission is prepared to preserve anonymity during review.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the implications of the results for uncertainty quantification in individualized decision-making settings. A positive societal impact is that the results may prevent overconfident use of distribution-free ITE prediction intervals under insufficient assumptions; a potential negative impact is that the results could be misinterpreted as discouraging all uncertainty quantification for causal effects, which we mitigate by clarifying the scope of the impossibility results and by presenting a positive result under additional structure.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, or sensitive data that would require safeguards for responsible release. The released materials, if any, consist only of anonymized simulation code for reproducing synthetic experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring

that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses existing open-source implementations of some baseline methods, and the original authors, repositories, versions, and licenses are credited in Appendix C. All such assets are used in accordance with their stated license terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases anonymized simulation code as supplementary material to reproduce the main experimental results. The code package includes documentation and instructions for generating the synthetic data and reproducing the reported figures, and no new dataset, pretrained model, or human-subject asset is introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments are based on synthetic data generated from the simulation designs described in Section 4 and Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. Therefore, IRB approval or equivalent human-subjects review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development, theoretical results, proofs, and experiments in this paper do not involve LLMs as an important, original, or non-standard component. Any use of LLMs, if applicable, was limited to writing, editing, or formatting assistance and did not affect the methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.