PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

Sensitivity analysis of individual treatment effects: A robust conformal inference approach

Ying Jin[a,1], Zhimei Ren[b,1], and Emmanuel J. Candès[a,c,2]

Contributed by Emmanuel J. Candès; received August 30, 2022; accepted December 16, 2022; reviewed by Victor Chernozhukov and Jing Lei

We propose a model-free framework for sensitivity analysis of individual treatment effects (ITEs), building upon ideas from conformal inference. For any unit, our procedure reports the Γ-value, a number which quantifies the minimum strength of confounding needed to explain away the evidence for ITE. Our approach rests on the reliable predictive inference of counterfactuals and ITEs in situations where the training data are confounded. Under the marginal sensitivity model of [Z. Tan, J. Am. Stat. Assoc. 101, 1619-1637 (2006)], we characterize the shift between the distribution of the observations and that of the counterfactuals. We first develop a general method for predictive inference of test samples from a shifted distribution; we then leverage this to construct covariate-dependent prediction sets for counterfactuals. No matter the value of the shift, these prediction sets (resp. approximately) achieve marginal coverage if the propensity score is known exactly (resp. estimated). We describe a distinct procedure also attaining coverage, however, conditional on the training data. In the latter case, we prove a sharpness result showing that for certain classes of prediction problems, the prediction intervals cannot possibly be tightened. We verify the validity and performance of the methods via simulation studies and apply them to analyze real datasets.

individual treatment effects | observational studies | sensitivity analysis | conformal inference | predictive inference

Significance

The individual treatment effect (ITE) describes the difference between an individual's outcome when receiving a treatment versus not. This difference may vary across individuals conditional on their characteristics. In observational studies, inference for ITEs can be invalid if one ignores unmeasured confounding factors that simultaneously influence the treatment assignment and the outcomes. We propose a framework to quantitatively understand the robustness of causal conclusions on ITEs against such potential confounding factors. This yields prediction bands, which come with rigorous uncertainty quantification tools. These tools apply regardless of the machine learning model employed to learn the treatment effect, however complicated, and regardless of the sample size.

Understanding the effect of a treatment is arguably one of the main research lines in causal inference. Over the past few decades, there has been a rich literature in identifying, estimating, and conducting inference on the mean value of causal effects; parameters of interest include the average treatment effect (ATE) or the conditional average treatment effect (CATE). These quantities, however, might fail to provide reliable uncertainty quantification for individual responses: The knowledge that a drug might be effective for a whole population "on average" does not imply that it is effective on a particular patient. Taking the intrinsic variability of the responses into account, inference on the individual treatment effect (ITE) may be better suited for reliable decision-making. To quantify the uncertainty in individual treatment effects, (1) offered a novel viewpoint. Rather than constructing confidence intervals for parameters—e.g., the ATE—they proposed designing prediction intervals for potential outcomes, namely, for counterfactuals and ITEs. Briefly, (1) constructed well-calibrated prediction intervals by building upon the conformal inference framework (2, 3). In that work, the typical mismatch between the counterfactuals and the observations due to the selection mechanism is resolved with the strong ignorability assumption (4); that is to say, the treatment assignment mechanism is independent of the potential outcomes conditional on a set of observed covariates. The strong ignorability assumption is automatically satisfied in randomized experiments and commonly used in the causal inference literature (4–6). In observational studies, however, the strong ignorability assumption is not testable (4, Chapter 12) and hard to justify in general. In practice, failing to account for possible confounding in observational data can yield misleading conclusions (7–9).

A. Γ-Values. In this paper, we seek to understand the robustness of causal conclusions on ITEs against potential unmeasured confounding. To this end, the procedure we propose starts from a sequence of hypothesized confounding strengths whose meaning will be made clear shortly; for each hypothesized strength, a prediction interval is constructed for the ITE. The procedure then screens the prediction intervals and, for each unit, reports the smallest confounding strength with which the prediction interval contains zero: we call this the Γ-value. Informally, the Γ-value describes the strength of unmeasured confounding necessary to explain away the predicted effect.

Reviewers: V.C., Massachusetts Institute of Technology; and J.L., Carnegie Mellon University.

The authors declare no competing interest.

[1]Y.J. and Z.R. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: candes@stanford.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2214889120/-/DCSupplemental.

Published February 2, 2023.

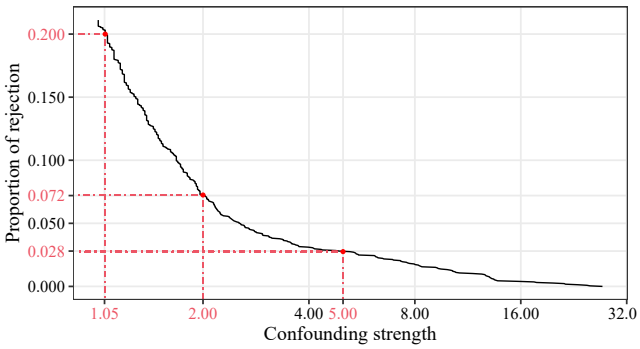

Fig. 1. Proportion of test samples in the real dataset from *Real Data Analysis* identified as positive ITE at each hypothesized confounding strength. The confidence level is set to $1 - \alpha = 0.9$.

Imagine we want to test for a positive ITE on a treated unit. For a range of hypothesized confounding strengths, our procedure constructs one-sided prediction intervals for the ITE at the $1 - \alpha = 0.9$ confidence level. The Γ-value is then the smallest confounding strength with which the lower bound of the prediction interval is smaller than zero. Fig. 1 shows the survival function of the Γ-values calculated on a real dataset measuring the academic performance of students subject or not to mindset interventions. (All the details are in *Real Data Analysis*). We see that 20% of the students have Γ-values greater than 1.05; roughly speaking, for these students, the confounding strength needs to be as large as 1.05 to explain away the evidence for positive ITEs. We also see that 7.20% of the students have Γ-values greater than 2 and some students have Γ-values as large as 5, showing strong evidence for positive ITEs.

Formally, the Γ-value can be used to draw conclusions on the ITE with a prespecified confounding strength. For example, if we believe that the confounding strength is no larger than 2, we can claim that an individual has positive ITE as long as its Γ-value is greater than 2. Our method guarantees that if the magnitude of confounding is at most 2, the probability of incorrectly "classifying" a unit as having a positive ITE is at most $\alpha = 0.1$.

Fig. 2 plots the Γ-values as a function of the achievement levels of the schools the students belong to. Once more, our procedure provides valid inference on a single ITE. This means that if the strength of confounding is at most 2, the chance that an individual with a negative ITE has a Γ-value larger than 2 is at most α. Taking a step further, one might be interested in the inference on a set of selected units (e.g., the red points in Fig. 2), for which evidence on multiple ITEs needs to be combined.

B. Problem Setup. Throughout, we work under the potential outcome framework (4, 10). Let $X \in \mathcal{X}$ denote the observed covariates, $T \in \{0, 1\}$ the assigned treatment, and $Y(1), Y(0) \in \mathbb{R}$ the potential outcomes. We assume that there is an unobserved confounder $U \in \mathcal{U}$ satisfying

$$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \,\big|\, X, U, \qquad [1]$$

in which U can be a random vector. As pointed out by ref. 11, a confounder satisfying Eq. **1** always exists since one can take $U = (Y(1), Y(0))$. In contrast, the strong ignorability assumption states

$$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \,\big|\, X. \qquad [2]$$

As well known, the latter assumes that we have measured sufficiently many features so that the potential outcomes are independent of the treatment conditional on X.

Suppose i.i.d. samples $\{(X_i, U_i, T_i, Y_i(0), Y_i(1))\}_{i\in\mathcal{D}}$ are from some distribution $\mathbb{P}$. Under the commonly used stable unit treatment value assumption (SUTVA) (e.g., refs. 4, 5, 12, 13), we observe $Y_i = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$; our observations are the triples $(X_i, T_i, Y_i)_{i\in\mathcal{D}}$.

Without further assumptions, the potential outcomes and the treatment assignment mechanism can arbitrarily depend on U, making the estimation of treatment effects impossible. For example, imagine we would like to assess the effect of a drug on patients. We are interested in $Y(1)$ (e.g., the survival time of the patient if the drug is taken) and have available observational data recording the treatment assignment T and outcome Y. Consider a confounded setting, where the drug is assigned to patients based on an undocumented factor U, namely, the patient's condition when admitted to the hospital, so that only those in critical condition get treated. Since U is highly correlated with $Y(1)$, the survival times of treated patients will likely be smaller than those in the whole population (which is our inferential target), making the task of identifying the effectiveness of the drug extremely difficult. As such, we shall work with confounders that have only limited effect on the treatment assignment mechanism; the concept of "limited effect" is formalized by the sensitivity models introduced below.

C. Sensitivity Models. A sensitivity model characterizes the degree to which the data distribution violates the strong ignorability assumption. There has been a rich literature in designing different types of sensitivity models (e.g., 14 and the references therein). In this paper, we work under the marginal sensitivity model (14, 15) on the unidentifiable superpopulation, characterized by the following marginal Γ-selection condition:

Definition 1: [Marginal Γ-selection] A distribution $\mathbb{P}$ over $(X, U, T, Y(0), Y(1))$ satisfies the marginal Γ-selection condition if for $\mathbb{P}$-almost all $x \in \mathcal{X}$ and $u \in \mathcal{U}$,

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(T = 1 \,|\, X = x, U = u)\mathbb{P}(T = 0 \,|\, X = x)}{\mathbb{P}(T = 0 \,|\, X = x, U = u)\mathbb{P}(T = 1 \,|\, X = x)} \le \Gamma. \qquad [3]$$

Intuitively, Γ bounds the relative distortion in the odds of receiving treatment versus control created by the unobserved confounders. Taking $\Gamma = 1$ is equivalent to the strong ignorability condition Eq. **2**, that is, no unmeasured confounding. Cornfield's seminal discussion (16) on the causal effect of smoking on lung cancer regards $\Gamma = 9$ as an unlikely confounding strength. Recent works on sensitivity analysis often hypothesize $\Gamma \in [1, 5]$. We have seen a value less than 3 to be sufficient to invalidate the conclusion on a nonzero average treatment effect (14). In practice, a plausible range for Γ often relies on domain knowledge.

The above marginal sensitivity model is closely related to Rosenbaum's sensitivity model (17) and its generalizations (11). As pointed out by (14, Proposition 7.1), a distribution $\mathbb{P}$ over $(X, U, T, Y(0), Y(1))$ satisfying their sensitivity models must also satisfy Definition 1. Our results thus also apply to the models considered in these works.

In the following, we let $\mathbb{P}^{\text{sup}}$ denote the unknown superpopulation over $(X, U, T, Y(0), Y(1))$ that generates the partial observations $\mathcal{D}$. For any $\Gamma \ge 1$, $\mathcal{P}(\Gamma)$ is the set of superpopulations that satisfy the marginal Γ-selection condition.

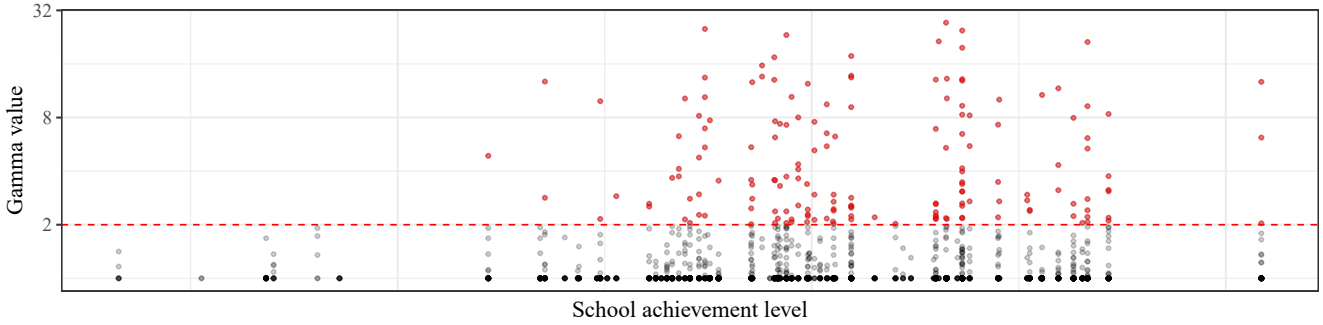

Fig. 2. Γ-values versus school achievement level for test samples. The red points correspond to the students whose Γ-values are greater than 2.

D. Prediction Intervals for Counterfactuals. The crux of our approach is to construct reliable prediction intervals for counterfactuals when the observations satisfy the marginal sensitivity model. In *Robust Weighted Conformal Inference*, we propose a generic robust weighted conformal procedure, which is applied to counterfactual prediction in *Counterfactual Inference with Confounding*. Suppose we are interested in $Y(1)$ and $(X_{n+1}, Y_{n+1}(1))$ is a test sample from the superpopulation (the results apply to other types of counterfactuals as well). Given a nominal level $1-\alpha$ and a fixed confounding level $\Gamma \geq 1$, the prediction interval $\widehat{C}(X_{n+1}, \Gamma)$ we construct from the confounded data ensures

$$\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}(X_{n+1}, \Gamma)\big) \geq 1-\alpha-\widehat{\Delta},$$

as long as $\mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma)$; here, the probability is over the confounded observations $(X_i, T_i, Y_i)_{i\in\mathcal{D}}$ and the test sample $(X_{n+1}, Y_{n+1}(1))$. The error term $\widehat{\Delta}=0$ if the propensity score $e(x)=\mathbb{P}(T=1\,|\,X=x)$ of the observed data is known exactly and otherwise depends on the estimation of the propensity score.

In practice, researchers may want to control the risk of falsely rejecting a hypothesis on an individual treatment effect given the data at hand, $\mathcal{D}$. *PAC-Type Robust Conformal Inference* offers a sister procedure with probably approximately correct (PAC)-type guarantee. Once more, suppose $\mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma)$. Then, given any $\delta, \alpha > 0$ and any $\Gamma \geq 1$, we can construct a prediction interval $\widehat{C}(X_{n+1}, \Gamma)$ such that

$$\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}(X_{n+1}, \Gamma)\,|\,\mathcal{D}\big) \geq 1-\alpha-\widehat{\Delta},$$

holds with probability at least $1-\delta$ over the randomness of $\mathcal{D}$. As before, the error term $\widehat{\Delta}$ depends on the estimation of the propensity score $e(x)$. Since any distribution $\mathbb{P}$ satisfying the Γ-selection condition must also satisfy the marginal Γ-selection condition, our methods also provide valid prediction intervals for counterfactuals under Rosenbaum's sensitivity model.

E. Related Work. The idea of sensitivity analysis dates back to ref. 16 who studied the causal effect of smoking on developing lung cancer. The authors concluded that if an unmeasured confounder—hormone in their example—were to rule out the causal association between smoking and lung cancer, it needed to be so strongly associated with smoking that no such factors could reasonably exist. The approach of ref. 16 requires that both the outcome and the confounder are binary and that there are no covariates. While (18, 19) used the same conditions later on, subsequent works substantially relaxed these assumptions. (20) proposed a sensitivity model to work with categorical covariates. Later, under Rosenbaum's sensitivity model, a series of works (17, 21–23) further extended sensitivity analysis to broader settings by studying samples with matching covariates (24–26) also considered unmeasured confounders with "limited" effect, but under different sensitivity models. More recently, (15) proposed the marginal sensitivity model, and (14) proposed a construction of bounds and confidence intervals for the ATE under this model. Their result was recently sharpened by (27). Bringing a distributionally robust optimization perspective to the sensitivity analysis problem, (11) studied the estimation of CATE under Rosenbaum's sensitivity model.

Our contribution is to provide robustness for the inference procedure against a proper level of confounding. This bears similarity with several works conducting "safe" policy evaluation and policy learning under certain sensitivity models (e.g., refs. 28 and 29). In contrast to the estimation and learning tasks, we provide well-calibrated uncertainty quantification for counterfactuals, which calls for a different set of techniques.

Another closely related line of work is conformal inference, which is the tool we employ (and improve) toward robust quantification of uncertainty. Conformal inference was pioneered and developed by Vladmir Vovk and his collaborators in a series of papers (e.g., refs. 2, 3, 30–33) and has connections to permutation tests (34). In recent years, the technique has been broadly used for establishing statistical guarantees for learning algorithms (e.g., refs. 35–39). In particular, (1) studied the counterfactual prediction problem with conformal inference tools under the strong ignorability condition. The PAC-type guarantee for conformal prediction sets is also studied in ref. 32 without distributional shifts.

In addition, our robust prediction perspective is related to ref. 40, which also studies the construction of robust prediction sets. That said, the setting considered there is substantially different from ours; we will expand on this in *Robust Weighted Conformal Inference*. (41) constructs PAC-type prediction sets under an identifiable covariate shift, with some robustness features. We provide in *PAC-Type Robust Conformal Inference* a robust PAC-type procedure as well; however, our methods apply to partially identifiable distributional shifts and are distinct from the rejection sampling strategy used in ref. 41.

Finally, we note that in an independent and concurrent paper, (42) also develops sensitivity analysis for the ITE under the marginal sensitivity model; they define the marginal sensitivity model without positing a latent confounder and provide an alternative derivation of our Lemma 2.1. They propose a robust weighted conformal inference procedure that is equivalent to Algorithm 1, while the analyses are complementary and offer different perspectives. Our current work additionally presents

a distinct algorithm achieving the PAC-type coverage and establishes the sharpness of our procedure in certain cases. We also interpret sensitivity analysis as a multiple testing procedure and prove that the type I error is simultaneously controlled over all superpopulations.

F. Outline of the Paper.

- *Robust Weighted Conformal Inference*, *Counterfactual Inference with Confounding*, and *PAC-Type Robust Conformal Inference* concern the development of robust counterfactual inference procedures. In *Robust Weighted Conformal Inference*, we develop a general robust weighted conformal inference procedure; we show in *Counterfactual Inference with Confounding* how to apply it to construct valid counterfactual prediction sets. We propose a distinct procedure in *PAC-Type Robust Conformal Inference* with PAC-type coverage, and establish a sharpness result.
- *Sensitivity Analysis of ITEs* expresses sensitivity analysis as a sequence of hypotheses testing problems and gives a statistical interpretation of the Γ-value. Simulation studies explain how the Γ-value relates to the true effect size and actual confounding level.
- *Real Data Analysis* evaluates the proposed method on a semisynthetic dataset to examine its validity and applicability. Finally, our sensitivity analysis framework is used to draw causal conclusions on a real dataset.

1. Robust Weighted Conformal Inference

We begin by considering a generic predictive inference problem under distributional shift. We will connect this to counterfactual inference under a marginal sensitivity model in *Counterfactual Inference with Confounding*.

Suppose we have i.i.d. training data $(X_i, Y_i)_{i\in\mathcal{D}}$ from some distribution $\mathbb{P}$ and an independent test sample (X_{n+1}, Y_{n+1}) from some possibly different distribution $\widetilde{\mathbb{P}}$. We consider a general setting where $\widetilde{\mathbb{P}}$ is "within bounded distance" from $\mathbb{P}$, in the sense that, for some fixed functions $\ell(\cdot)$ and $u(\cdot)$, it belongs to the identification set defined as

$$\mathcal{P}(\mathbb{P}, \ell, u) = \left\{ \widetilde{\mathbb{P}} \colon \ell(x) \le \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(x, y) \le u(x) \ \ \mathbb{P}\text{-a.s.} \right\}. \qquad [4]$$

The task is to provide a calibrated prediction interval $\widehat{C}(X_{n+1})$ for Y_{n+1}.

Eq. **4** identifies a class of distributional robustness problems. As we shall see later in *Bounding the Distributional Shift*, our model Eq. **4** is motivated by sensitivity analysis and is quite distinct from other models in the literature. For instance, (40) considers a setting in which the target distribution $\widetilde{\mathbb{P}}$ is assumed to be within an f-divergence ball with radius ρ centered around $\mathbb{P}$, so that the identification set is

$$\mathcal{Q}(\mathbb{P}, \rho) = \left\{ \widetilde{\mathbb{P}} : D_f\big(\widetilde{\mathbb{P}} \,\|\, \mathbb{P}\big) \le \rho \right\}. \qquad [5]$$

Instead of bounding the overall shift in Eq. **5**, the constraint in Eq. **4** actually allows freedom in the shift of X; to be sure, the set Eq. **4** can be small as long as $\ell(x)$ and $u(x)$ are close. For counterfactual prediction under the strong ignorability condition in ref. 1, the identification set Eq. **4** is a singleton even if $\mathbb{P}_X$ and $\widetilde{\mathbb{P}}_X$ are drastically different. This happens because in this case, $\ell(x) \equiv u(x)$, whereas Eq. **5** would require a large value of ρ to hold. More generally, when there is a (approximately) known large shift in the marginal distribution $\mathbb{P}_X$ but a relatively small shift in the conditional $\mathbb{P}_{Y\,|\,X}$, Eq. **4** provides a tighter range of the target distributions. Finally, the pointwise constraint in Eq. **4** (as opposed to the average form) makes it naturally compatible with the weighted conformal inference framework of ref. 43.

A. Warm-Up: Weighted Conformal Inference. Before introducing our method, it is best to start by a brief recap of the weighted (split) conformal inference procedure. Assume that the likelihood ratio $w(x, y) = \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(x, y)$ is known exactly. The dataset $\mathcal{D}$ is first randomly split into a training fold $\mathcal{D}_{\text{train}}$ of cardinality n_{train} and a calibration fold $\mathcal{D}_{\text{calib}}$ of cardinality n. We use $\mathcal{D}_{\text{train}}$ to train any nonconformity score function $V \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures how well (x, y) "conforms" to the calibration samples: the smaller $V(x, y)$, the better (x, y) conforms to the calibration samples; see e.g., ref. 44 for examples of nonconformity scores. We then define $V_i = V(X_i, Y_i)$ for $i \in \mathcal{D}_{\text{calib}}$. For any hypothetical value $(x, y) \in \mathcal{X} \times \mathcal{Y}$ of the new data point, we assign weights to the samples as

$$p_i^w(x, y) := \frac{w(X_i, Y_i)}{\sum_{j=1}^n w(X_j, Y_j) + w(x, y)}, \ i = 1, \ldots, n,$$

$$p_{n+1}^w(x, y) := \frac{w(x, y)}{\sum_{j=1}^n w(X_j, Y_j) + w(x, y)}.$$

For any random variable Z, define the quantile function as $\text{Quantile}(q, Z) = \inf\{z \colon \mathbb{P}(Z \le z) \ge q\}$, and let δ_a denote a point mass at a. Then, a level $(1-\alpha)$ prediction interval is given by

$$\widehat{C}(X_{n+1}) = \big\{ y : V(X_{n+1}, y) \le \widehat{V}_{1-\alpha}(y) \big\}, \qquad [6]$$

where $\widehat{V}_{1-\alpha}(y) = \text{Quantile}\big(1-\alpha, \ \sum_{i=1}^n p_i^w(X_{n+1}, y) \cdot \delta_{V_i} + p_{n+1}^w(X_{n+1}, y) \cdot \delta_\infty\big)$. The prediction interval Eq. **6** is shown by ref. 43 to obey $\widetilde{\mathbb{P}}(Y_{n+1} \in \widehat{C}(X_{n+1})) \ge 1-\alpha$, and it is computable when $w(x, y)$ is a known function of x only. In our context, $w(x, y)$ depends on x only when $\mathcal{P}(\mathbb{P}, \ell, u) = \{\widetilde{\mathbb{P}}\}$ is a singleton—this is exactly the case in ref. 1, where condition Eq. **2** is assumed.

B. Robust Weighted Conformal Inference Procedure. Now, suppose we have a pair of functions $\widehat{\ell}$ and $\widehat{u} \colon \mathcal{X} \to \mathbb{R}^+$ with $\widehat{\ell}(x) \le \widehat{u}(x)$ for all $x \in \mathcal{X}$. In general, we expect $\widehat{\ell}(x)$ (resp. $\widehat{u}(x)$) to serve as a pointwise lower (resp. upper) bound on the unknown likelihood ratio $w(x, y)$, although our theoretical guarantee does not depend on this.

Proceeding as before and denoting $\mathcal{D}_{\text{calib}} = \{1, \ldots, n\}$, we train a nonconformity score function $V \colon \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ on $\mathcal{D}_{\text{train}}$. We also allow $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$ to be obtained from $\mathcal{D}_{\text{train}}$. Let $[1], [2], \ldots, [n]$ be a permutation of $\{1, 2, \ldots, n\}$ such that $V_{[1]} \le V_{[2]} \le \cdots \le V_{[n]}$. Defining $\ell_i = \widehat{\ell}(X_i)$ and $u_i = \widehat{u}(X_i)$ for $1 \le i \le n$, and $u_{n+1} = \widehat{u}(X_{n+1})$, we construct the prediction interval

$$\widehat{C}(X_{n+1}) = \big\{ y : V(X_{n+1}, y) \le V_{[k^*]} \big\},$$

where $k^* = \min\{k : \widehat{F}(k) \ge 1-\alpha\}$ for

$$\widehat{F}(k) = \frac{\sum_{i=1}^k \ell_{[i]}}{\sum_{i=1}^k \ell_{[i]} + \sum_{i=k+1}^n u_{[i]} + u_{n+1}}. \qquad [7]$$

The thresholding function $\widehat{F}(k)$ in Eq. 7 is monotone in k; hence, a linear search suffices to find k^*. We summarize the procedure in Algorithm 1.

Algorithm 1: Robust conformal prediction

1: Input: Calibration data $\mathcal{D}_{\text{calib}}$, bounds $\widehat{\ell}(\cdot)$, $\widehat{u}(\cdot)$, nonconformity score function $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, test covariate x, target level $\alpha \in (0, 1)$.
2: For $i \in \mathcal{D}_{\text{calib}}$, compute $V_i = V(X_i, Y_i)$
3: For $i \in \mathcal{D}_{\text{calib}}$, compute $\ell_i = \widehat{\ell}(X_i)$ and $u_i = \widehat{u}(X_i)$.
4: Compute $u_{n+1} = \widehat{u}(x)$.
5: For each $1 \le k \le n$, compute $\widehat{F}(k)$ as in Eq. 7.
6: Compute $k^* = \min\{k\colon \widehat{F}(k) \ge 1 - \alpha\}$.
7: Output: Prediction set $\widehat{C}(x) = \{y\colon V(x, y) \le V_{[k^*]}\}$.

Remark 1.1. Writing $W_i = w(X_i, Y_i)$ for $1 \le i \le n$ and $W_{n+1}(y) = w(X_{n+1}, y)$, since the function $f(x, a) = x/(x + a)$ is increasing in $x > 0$ and decreasing in $a > 0$, we can check that $\widehat{V}_{1-\alpha}(y) = V_{[k^*(y)]}$, where $k^*(y) = \min\{k\colon F(k, y) \ge 1 - \alpha\}$ and

$$F(k, y) = \frac{\sum_{i=1}^{k} W_{[i]}}{\sum_{i=1}^{n} W_i + W_{n+1}(y)}. \qquad [8]$$

For each k, the threshold $\widehat{F}(k)$ in Eq. 7 is the solution to the following optimization problem

$$\begin{aligned} \text{minimize} \quad & \frac{\sum_{i=1}^{k} W_{[i]}}{\sum_{i=1}^{n} W_i + W_{n+1}} \\ \text{subject to} \quad & \widehat{\ell}(X_i) \le W_i \le \widehat{u}(X_i), \quad \forall\, i \in \mathcal{D}_{\text{calib}} \cup \{n+1\}, \end{aligned}$$

which seeks a lower bound on the unknown $F(k, y)$ in Eq. 8 if we believe $\widehat{\ell}(x) \le w(x, y) \le \widehat{u}(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Therefore, $\widehat{V}_{[k^*]}$ is a conservative estimate (upper bound) of $\widehat{V}_{1-\alpha}(y)$ in Eq. 6, and it is also the tightest upper bound one could obtain given the constraints $\widehat{\ell}(X_i) \le W_i \le \widehat{u}(X_i)$, $\forall\, i \in \mathcal{D}_{\text{calib}} \cup \{n+1\}$.

C. Theoretical Guarantee. To state the coverage guarantee of Algorithm 1, we start with some notations. We write $a_+ = \max(a, 0)$ and $a_- = \max(-a, 0)$ for any $a \in \mathbb{R}$. For any random variable U, define $\|U\|_r = (\mathbb{E}[|U|^r])^{1/r}$ as the L_r norm for any $r \ge 1$, and $\|U\|_\infty = \mathbb{E}[\operatorname{ess\,sup} |U|]$ as the L_∞ norm. Throughout the paper, all the statements are conditional on $\mathcal{D}_{\text{train}}$, so that $\widehat{\ell}(\cdot)$, $\widehat{u}(\cdot)$ and $V(\cdot, \cdot)$ can be viewed as fixed functions.

Theorem 1.2. *Assume that* $(X_i, Y_i)_{i \in \mathcal{D}_{\text{calib}}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, *and the independent test point* $(X_{n+1}, Y_{n+1}) \sim \mathbb{P}$ *has likelihood ratio* $w(x, y) = \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}(x, y)$. *Then, for any target level* $\alpha \in (0, 1)$, *the output of Algorithm 1 satisfies*

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \ge 1 - \alpha - \widehat{\Delta},$$

where the probability is over $\mathcal{D}_{\text{calib}}$ *and* (X_{n+1}, Y_{n+1}), *and*

$$\begin{aligned} \widehat{\Delta} = \big\|1/\widehat{\ell}(X)\big\|_q \Big(& \big\|(\widehat{\ell}(X) - w(X, Y))_+\big\|_p \\ & + \big\|(\widehat{u}(X) - w(X, Y))_-\big\|_p \\ & + \frac{1}{n}\big\|w(X, Y)^{1/p} \cdot (\widehat{u}(X) - w(X, Y))_-\big\|_p \Big). \end{aligned} \qquad [9]$$

Given $q \ge 1$, p *is chosen such that* $1/p + 1/q = 1$ *with the convention that* $p = \infty$ *for* $q = 1$. *The expectation in* L_p, L_q *is taken over an independent sample* $(X, Y) \sim \mathbb{P}$.

Clearly, if $\widehat{\ell}(X) \le w(X, Y) \le \widehat{u}(X)$ *a.s.,* $\widehat{\Delta} = 0$ *and*

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \ge 1 - \alpha.$$

The proof of Theorem 1.2 is deferred to *SI Appendix*, Appendix 3.B. Almost sure bounds on the likelihood ratio imply exact coverage. Otherwise, coverage is off by at most $\widehat{\Delta}$. First, $\|1/\widehat{\ell}(X)\|_q$ is bounded when $\widehat{\ell}(\cdot)$ is bounded away from zero. Second, the remaining terms (between brackets) are small if $\widehat{\ell}(X)$ (resp. $\widehat{u}(X)$) is below (resp. above) the likelihood ratio most of the time. In particular, for all target distributions within Eq. 4, if $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$ are estimators for $\ell(\cdot)$ and $u(\cdot)$, one additive term can be decomposed as

$$\begin{aligned} & \big\|(\widehat{\ell}(X) - w(X, Y))_+\big\|_p \\ & = \Big\| \big\{ \underbrace{[\widehat{\ell}(X) - \ell(X)]}_{\text{estimation error}} - \underbrace{[w(X, Y) - \ell(X)]}_{\text{population gap}} \big\}_+ \Big\|_p, \end{aligned} \qquad [10]$$

which is small as long as the estimation error does not exceed the population gap most of the time. Our numerical experiments demonstrate that $\widehat{\Delta}$ is reasonably small even with nonnegligible estimation errors.

Remark 1.3. The miscoverage bound in Theorem 1.2 holds for any distributional shift and any postulated bounds $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$. The results also hold if the bounds take the general form $\widehat{\ell}(x, y)$ and $\widehat{u}(x, y)$. Therefore, our procedure may be of use in a very broad range of settings.

2. Counterfactual Inference with Confounding

With the generic methodology in place, we return to counterfactual inference under unmeasured confounding. We shall characterize distributional shifts of interest, translate the target distribution into a set like Eq. 4, and show how Algorithm 1 can be applied.

Recall that we have a partially revealed and possibly confounded dataset $(X_i, T_i, Y_i)_{i \in \mathcal{D}}$ generated by an unknown super population $\mathbb{P}^{\text{sup}}$. As an example, we consider an independent new unit from $\mathbb{P}^{\text{sup}}$ with only the covariate X_{n+1} but no outcomes observed. We would like to construct a prediction interval that covers $Y_{n+1}(1)$ with probability at least $1 - \alpha$; the probability is over the randomness of the training sample $\mathcal{D}$ and the new unit. The main challenge is that the distribution of the new unit may differ from that of the observations we have access to; that is, the training distribution is $\mathbb{P}_{\text{train}} = \mathbb{P}_{X, Y(1) \mid T=1}$, while the target distribution is $\mathbb{P}_{\text{target}} = \mathbb{P}_{X, Y(1)}$.

A. Bounding the Distributional Shift. In the previous example, for a new unit from the superpopulation, the likelihood ratio

takes the form $w(x,y) = \frac{d\mathbb{P}_{X,Y(1)}}{d\mathbb{P}_{X,Y(1)\,|\,T=1}}(x,y)$. A key observation in ref. 1 is that, under the strong ignorability condition Eq. **2**, $w(x,y)$ is an identifiable function of x, i.e., the target distribution can be identified from the observed (training) distribution with a covariate shift. This fact is used to construct prediction intervals for counterfactuals with finite-sample guarantees by leveraging the weighted conformal inference procedure.

In the presence of unmeasured confounding, $w(x,y)$ can no longer be expressed as a function of x. However, under the marginal Γ-selection condition Eq. **3**, $w(x,y)$ can be bounded from above and below by functions of x; that is, the unknown target distribution falls within a set of the form Eq. **4**. The following lemma is a key ingredient for establishing the boundedness result.

Lemma 2.1. *Suppose* $\mathbb{P}$ *over* $(X, U, T, Y(0), Y(1))$ *satisfies the marginal* Γ*-selection condition. Then, for any* $t \in \{0,1\}$*, it holds for* $\mathbb{P}$*-almost all* $x \in \mathcal{X}$*,* $y \in \mathcal{Y}$ *that*

$$\frac{1}{\Gamma} \le \frac{d\mathbb{P}_{Y(t)\,|\,X,T=t}}{d\mathbb{P}_{Y(t)\,|\,X,T=1-t}}(x,y) \le \Gamma.$$

The proof of Lemma 2.1 is in *SI Appendix*, Appendix 3.A. Returning to $w(x,y)$, letting $\bar{p} = \mathbb{P}(T=1)$, by Bayes' rule, we have

$$\frac{d\mathbb{P}_{X,Y(1)}}{d\mathbb{P}_{X,Y(1)\,|\,T=1}} = \bar{p}\cdot\left(1 + \frac{d\mathbb{P}_{Y(1)\,|\,X,T=1}}{d\mathbb{P}_{Y(1)\,|\,X,T=0}}\frac{1-e(X)}{e(X)}\right). \quad [11]$$

Applying Lemma 2.1 to Eq. **11**, we could then bound the likelihood ratio $w(x,y)$ by functions of the covariate x.

The above reasoning broadly applies to other types of inferential targets. For example, for a new treated unit (i.e., given that $T_{n+1}=1$), we might consider average treatment effect on the treated (ATT)-type inference on $Y(1)$; that is, to construct a prediction interval such that $\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}(X_{n+1}) \,|\, T_{n+1}=1\big) \ge 1-\alpha$. In this case, the likelihood ratio is simply $w(x,y) = \frac{d\mathbb{P}_{X,Y(1)\,|\,T=1}}{d\mathbb{P}_{X,Y(1)\,|\,T=1}}(x,y) = 1$. Alternatively, for a new control unit (i.e., given $T_{n+1}=0$), we might be interested in the average treatment effect on the control (ATC)-type inference on $Y(1)$; that is, to construct $\widehat{C}(X_{n+1})$ such that $\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}(X_{n+1}) \,|\, T=0\big) \ge 1-\alpha$. In this case, the likelihood ratio is $\frac{d\mathbb{P}_{X,Y(1)\,|\,T=0}}{d\mathbb{P}_{X,Y(1)\,|\,T=1}}$, bounded by $\ell(x) = \frac{\bar{p}}{1-\bar{p}}\cdot\frac{1-e(x)}{\Gamma\cdot e(x)}$, $u(x) = \frac{\bar{p}}{1-\bar{p}}\cdot\frac{1-e(x)}{\Gamma^{-1}\cdot e(x)}$. More generally, one might also wish to conduct inference on a unit from a different population, i.e., the target distribution admits a different distribution of covariates, whereas the joint distribution of $(Y(0), Y(1), U, T)$ given X stays invariant. To be specific, we assume that the training distribution is $\mathbb{P}_{X,U,T,Y(0),Y(1)} = \mathbb{P}_X \times \mathbb{P}_{U,T,Y(0),Y(1)\,|\,X}$, while the target distribution is $\mathbb{P}_{X,U,T,Y(0),Y(1)} = \mathbb{Q}_X \times \mathbb{P}_{U,T,Y(0),Y(1)\,|\,X}$. The corresponding upper and lower bounds on the likelihood ratio are $\ell(x) = \bar{p}\cdot\frac{d\mathbb{Q}_X}{d\mathbb{P}_X}(x)\cdot\big(1+\frac{1-e(x)}{\Gamma\cdot e(x)}\big)$, $u(x) = \bar{p}\cdot\frac{d\mathbb{Q}_X}{d\mathbb{P}_X}(x)\cdot\big(1+\frac{1-e(x)}{\Gamma^{-1}\cdot e(x)}\big)$. All the above types of inferential targets can also be applied to $Y(0)$ following the same arguments. We summarize the bounds for various inferential targets in *SI Appendix*, Table S1 of Appendix 1. which recovers Table 1 in ref. 1 when $\Gamma = 1$.

B. Robust Counterfactual Inference. To apply Algorithm 1 to counterfactual prediction with a prespecified confounding level Γ, we take the upper and lower bounds as presented in *SI Appendix*, Table S1. For instance, for ATE-type predictive inference for $Y(1)$, the likelihood ratio $w(x,y)$ is bounded by $\ell(x)$ and $u(x)$, which take the form:

$$\ell(x) = \bar{p}\cdot\left(1+\frac{1-e(x)}{\Gamma\cdot e(x)}\right),\quad u(x) = \bar{p}\cdot\left(1+\frac{1-e(x)}{\Gamma^{-1}\cdot e(x)}\right).$$

We then construct $\widehat{\ell}(x)$ and $\widehat{u}(x)$ by plugging in an estimator $\widehat{e}(x)$ of $e(x)$ trained on $\mathcal{D}_{\text{train}}$. Taking $\Gamma = 1$, our procedure recovers the approach of ref. 1, where $\widehat{\ell}(x) = \widehat{u}(x)$ and both equal the likelihood ratio function; our finite-sample bound $\widehat{\Delta}$ in Eq. **9** reduces to a bound similar to (1, Theorem 3) (the forms of the two bounds are slightly different, hence not directly comparable in general).

The finite-sample guarantee in Theorem 1.2 shows that the accuracy of $\widehat{e}(x)$ itself (hence that of $\widehat{\ell}(x)$ and $\widehat{u}(x)$) may not matter much for valid coverage; what matters is whether or not $\widehat{\ell}(x)$ is below $w(x,y)$ and $\widehat{u}(x)$ above $w(x,y)$. In our simulation studies, we empirically evaluate $\|\widehat{\ell}(X)-\ell(X)\|_1$, $\|\widehat{u}(X)-u(x)\|_1$ and $\widehat{\Delta}$ (Fig. 4). Even if $\|\widehat{\ell}(X)-\ell(X)\|_1$ and $\|\widehat{u}(X)-u(X)\|_1$ can be large (especially for large Γ), the gap $\widehat{\Delta}$ remains small.

C. Numerical Experiments. We illustrate the performance of the novel procedure in a simulation setting similar to that in ref. 11. Given a sample size $n_{\text{train}} = n_{\text{calib}} \in \{500, 2000, 5000\}$ and a covariate dimension $p \in \{4, 20\}$, we generate the covariates and unobserved confounders as $X \sim \text{Unif}[0,1]^p$, $U\,|\,X \sim N(0, 1+\frac{1}{2}\cdot(2.5X_1)^2)$. The counterfactual of interest is $Y(1)$ generated as $Y(1) = \beta^\top X + U$, where $\beta = (-0.531, 0.126, -0.312, 0.018, 0, \ldots, 0)^\top \in \mathbb{R}^p$. With i.i.d. data (X_i, Y_i, U_i) generated from the fixed superpopulation, T_i with different confounding levels $\Gamma \in \{1, 1.5, 2, 2.5, 3, 5\}$ satisfying Eq. **3**. Specifically, we design the propensity scores as $e(x) = \text{logit}(\beta^\top x)$ and

$$e(x,u) = a(x)\,\mathbb{1}\{|u| > t(x)\} + b(x)\,\mathbb{1}\{|u| \le t(x)\}, \quad [12]$$

for the same $\beta \in \mathbb{R}^p$. Above, $a(x) = \frac{e(x)}{e(x)+\Gamma(1-e(x))}$, $b(x) = \frac{e(x)}{e(x)+(1-e(x))/\Gamma}$ are the lower and upper bounds on $e(x,u)$ under the marginal Γ-selection model. The threshold $t(x)$ is designed to ensure $\mathbb{E}[e(X,U)\,|\,X] = e(X)$. The training data $\mathcal{D}$ are $\{(X_i, Y_i(1)\}$ for those $T_i = 1$. By Eq. **12**, the setting is designed to be adversarial so as to show the performance of our method in a nearly worst case.

For each configuration of $(n_{\text{calib}}, p, \Gamma, \alpha)$, we run Algorithm 1 with ground truth $\ell(\cdot)$, $u(\cdot)$ and with estimated $\widehat{\ell}(\cdot)$, $\widehat{u}(\cdot)$. We fit the propensity score $\widehat{e}(x)$ on $\mathcal{D}_{\text{train}}$ with `grf` R-package, and set $\widehat{\ell}(x) = \widehat{p}_1(1+(1-\widehat{e}(x))/(\Gamma\cdot\widehat{e}(x)))$, $\widehat{u}(x) = \widehat{p}_1(1+\Gamma\cdot(1-\widehat{e}(x))/\widehat{e}(x))$ with $\widehat{p}_1 = \frac{1}{n_{\text{train}}}\sum_{i=1}^{n_{\text{train}}} T_i$. We compute the average coverage on the test sample over $N = 1{,}000$ independent runs.

The proposed approaches work for any nonconformity score function and any training method. In our experiments, we follow the conformalized quantile regression algorithm (CQR) (45) to compute the nonconformity score: $\mathcal{D}_{\text{train}}$ is used to train a conditional quantile function $\widehat{q}(x,\beta)$ for $Y(1)$ given X by quantile random forests (46). Then, for a target level $\alpha \in (0,1)$, we define $V(x,y) = \max\{\widehat{q}(x,\alpha/2) - y,\ y - \widehat{q}(x,1-\alpha/2)\}$. We run the procedures for a target $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$ with the corresponding nonconformity scores.

The empirical coverage with estimated bounds is in Fig. 3; its counterpart with ground truth in *SI Appendix*, Fig. S1 in Appendix 4.A shows quite similar performance.

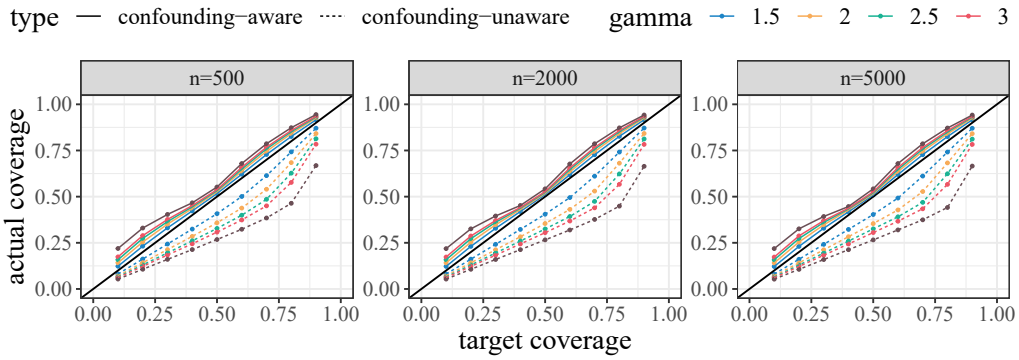

Fig. 3. Empirical coverage of Algorithm 1 with estimated bounds. Each plot corresponds to a sample size $n = n_{\text{calib}}$ while fixing $p = 20$. Within a plot, each line represents a confounding level Γ. The solid lines are for Algorithm 1. The dashed lines assume no confounding and are shown for comparison.

In Fig. 3, the solid lines are all above the 45°-line, showing the validity of our procedure over the whole spectrum of sample sizes. We also see that confounding must be taken into consideration to reach valid counterfactual conclusions. The coverage is close to the target, showing the sharpness of the prediction interval in this setting.

As illustrated in Eq. **10**, the gap between $\ell(x)$ and $w(x, y)$ (and that between $w(x, y)$ and $u(x)$) provides some buffer for the estimation error in $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$. We numerically evaluate the gap $\widehat{\Delta}$ with $(q, p) = (\infty, 1)$ as well as the estimation error $\|\widehat{\ell}(X) - \ell(X)\|_1$ and $\|\widehat{u}(X) - u(X)\|_1$ in Fig. 4. Although the estimation errors can sometimes be large, the realized gap $\widehat{\Delta}$ is often very close to zero.

3. PAC-Type Robust Conformal Inference

We describe a distinct procedure to construct robust prediction sets with guaranteed coverage on the test sample conditional on the training data. Such guarantee has been considered by refs. 47 and 48 and might be appealing to the practitioners—it ensures that unless one gets really unlucky with the training set $\mathcal{D}$, the anticipated coverage of the prediction set on the test sample conditional on $\mathcal{D}$ achieves the desired level. This procedure is based on the worst case distribution of nonconformity scores, taking a different perspective from the previous one.

A. The Procedure. We state our approach in the generic setting, starting with some basic notations. Throughout, we follow the sample splitting routine as before, and all statements are conditional on $\mathcal{D}_{\text{train}}$. Suppose we have a pair of functions $\widehat{\ell}$ and $\widehat{u}$: $\mathcal{X} \to \mathbb{R}^+$ with $\widehat{\ell}(x) \le \widehat{u}(x)$ for all $x \in \mathcal{X}$. Again, we expect them to bound the unknown likelihood ratio in a pointwise fashion, although this is not required for getting theoretical guarantees.

Our approach is based on a general nondecreasing function $G(\cdot)\colon \mathbb{R} \to [0, 1]$. We expect $G(\cdot)$ to serve as a conservative envelope function for the unknown target distribution of the nonconformity score, i.e., $G(t) \le \widetilde{\mathbb{P}}(V(X, Y) \le t)$ for all $t \in \mathbb{R}$. For now, we construct it as

$$G(t) = \max\Big\{\mathbb{E}[\mathbb{1}_{\{V(X,Y)\le t\}}\,\widehat{\ell}(X)],\ 1 - \mathbb{E}[\mathbb{1}_{\{V(X,Y)>t\}}\,\widehat{u}(X)]\Big\}, \qquad [13]$$

where $\mathbb{E}$ is with respect to an independent copy $(X, Y) \sim \mathbb{P}$. It lower bounds $\widetilde{\mathbb{P}}(V(X, Y) \le t)$ when $\widehat{\ell}(x)$ and $\widehat{u}(x)$ are lower and upper bounds on $w(x, y)$.

Given a constant $\delta \in (0, 1)$, suppose we can construct a nondecreasing confidence lower bound $\widehat{G}_n(\cdot)$ for $G(\cdot)$ such that for any fixed $t \in \mathbb{R}$, it holds that

$$\mathbb{P}_{\mathcal{D}}\big(\widehat{G}_n(t) \le G(t)\big) \ge 1 - \delta, \qquad [14]$$

where $\mathbb{P}_{\mathcal{D}}$ is taken with respect to $\mathcal{D}_{\text{calib}}$. We then define the prediction interval as

$$\widehat{C}(X_{n+1}) = \big\{y : V(X_{n+1}, y) \le \inf\{t : \widehat{G}_n(t) \ge 1 - \alpha\}\big\}.$$

The procedure is summarized in Algorithm 2.

Algorithm 2: PAC robust conformal prediction

1: Input: Calibration data $\mathcal{D}_{\text{calib}}$, bounds $\widehat{\ell}(\cdot)$, $\widehat{u}(\cdot)$, nonconformity score $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, test covariate x, target level $\alpha \in (0, 1)$, confidence level $\delta \in (0, 1)$.
2: Construct the conservative envelope distribution function $\widehat{G}_n(t)$ for $t \in \mathbb{R}$ so that Eq. **14** holds.
3: Compute $\widehat{v} = \inf\{t\colon \widehat{G}_n(t) \ge 1 - \alpha\}$.
4: Output: Prediction set $\widehat{C}(x) = \{y\colon V(x, y) \le \widehat{v}\}$.

Construction of estimators. Assume that $\|\widehat{\ell}\|_\infty,\ \|\widehat{u}\|_\infty \le M$ for some constant $M > 0$, then we can construct $\widehat{G}_n(t)$ via $\widehat{G}_n(t) = \max\Big\{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{V_i \le t\}}\,\widehat{\ell}(X_i),\ 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{V_i > t\}}\,\widehat{u}(X_i)\Big\} - M\sqrt{\frac{\log(2/\delta)}{2n}}$. Hoeffding's inequality implies that Eq. **14** holds for any fixed $t \in \mathbb{R}$. In our numerical experiments, we construct $\widehat{G}_n(\cdot)$ via the Waudby-Smith–Ramdas bound (49) and defer details to Proposition 2.1 in *SI Appendix*, Appendix 2.A. When $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$ are obtained from $\mathcal{D}_{\text{train}}$, the upper bounds on $\|\widehat{\ell}\|_\infty$ and $\|\widehat{u}\|_\infty$ become readily available.

B. Theoretical Guarantee.

Theorem 3.1. *Assume* $(X_i, Y_i)_{i\in\mathcal{D}_{\text{calib}}} \overset{i.i.d.}{\sim} \mathbb{P}$ *and the independent test point* $(X_{n+1}, Y_{n+1}) \sim \widetilde{\mathbb{P}}$ *has likelihood ratio* $w(x, y) = \frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}(x, y)$. *Fix a target level* $\alpha \in (0, 1)$ *and a confidence level* $\delta \in (0, 1)$. *Suppose* $\widehat{G}_n(\cdot)$ *satisfies* Eq. **14** *for* $G(\cdot)$ *in* Eq. **13**, *then the output of Algorithm 2 obeys*

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\text{calib}}\big) \ge 1 - \alpha - \widehat{\Delta}$$

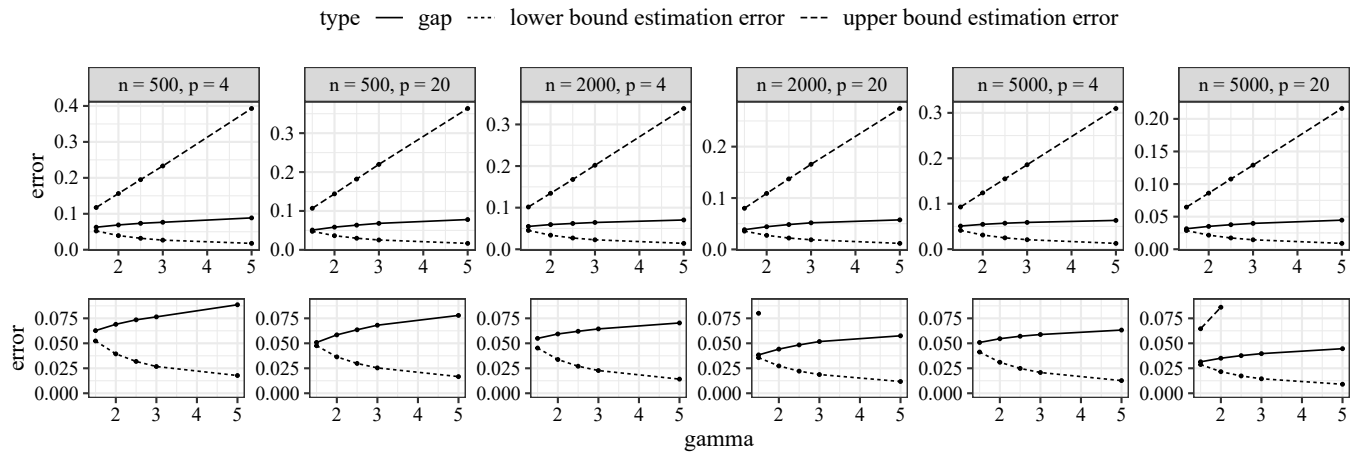

Fig. 4. Empirical gap and estimation errors. The plots in the second row zoom in on the gaps. Each plot corresponds to a sample size $n = n_{\text{calib}}$ and a dimension p. The long-dashed lines are $\|\widehat{u}(X) - u(X)\|_1$, the short-dashed lines are $\|\widehat{\ell}(X) - \ell(X)\|_1$, and the solid lines are $\widehat{\Delta}$ defined in Theorem 1.2.

with probability at least $1 - \delta$ *over* $\mathcal{D}_{\text{calib}}$*, and*

$$\widehat{\Delta} = \max\left\{\mathbb{E}\big[(\widehat{\ell}(X) - w(X,Y))_+\big],\ \mathbb{E}\big[(\widehat{u}(X) - w(X,Y))_-\big]\right\}. \quad [15]$$

The expectations are over an independent copy $(X, Y) \sim \mathbb{P}$*. If* $\widehat{\ell}(X) \le w(X,Y) \le \widehat{u}(X)$ *a.s., then* $\widehat{\Delta} = 0$ *and*

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\text{calib}}\big) \ge 1 - \alpha.$$

The proof of Theorem 3.1 is deferred to *SI Appendix*, Appendix 3.C. Almost sure bounds on the likelihood ratio yields exact coverage. Otherwise, coverage is off by at most $\widehat{\Delta}$. Again, $\widehat{\Delta}$ is small if $\widehat{\ell}(X)$ (resp. $\widehat{u}(X)$) is below (resp. above) $w(X, Y)$ most of the time. This is demonstrated in the simulation results presented in Fig. 6.

Just as before, the miscoverage bound in Theorem 3.1 holds for any distributional shift and any inputs $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$. The conclusion also applies to inputs of the form $\widehat{\ell}(x, y)$ and $\widehat{u}(x, y)$ without modification.

Application to counterfactual prediction. To apply Algorithm 2 to counterfactual prediction under a prespecified confounding level Γ, We plug in the estimated $\widehat{e}(x)$ to obtain $\widehat{\ell}(x)$ and $\widehat{u}(x)$ according to *SI Appendix*, Table S1. As before, $\widehat{\Delta}$ can be small as long as $\widehat{\ell}(x)$ (resp. $u(x)$) falls below (resp. above) $w(x)$ most of the time, even if $\widehat{e}(x)$ has a nonnegligible estimator error.

C. Sharpness. One may also be concerned with the sharpness of the method—indeed, one can always construct a valid but arbitrarily conservative prediction interval. Ideally, we desire a valid prediction set whose coverage is not too much larger than the prescribed level.

We take a close look at the sharpness of Algorithm 2 by identifying the worst-case distributional shift. We study the sharpness of our method in two problems: robust prediction and counterfactual inference under unmeasured confounding. They are treated in the same way when developing the validity results, but they actually have distinct identification sets and sharpness properties.

Given an identification set $\mathcal{P}$ of the unknown target distribution and viewing the score function V as fixed, we define the worst-case c.d.f. of $V(X, Y)$ under $\mathcal{P}$ as

$$F(t; \mathcal{P}) := \inf_{\widetilde{\mathbb{P}} \in \mathcal{P}} f(t, \widetilde{\mathbb{P}}) = \inf_{\widetilde{\mathbb{P}} \in \mathcal{P}} \widetilde{\mathbb{P}}\big(V(X,Y) \le t\big). \quad [16]$$

The probability is over an independent copy $(X, Y) \sim \mathbb{P}$.

To remove the nuisance in estimation, we consider the asymptotic regime $\widehat{G}_n(t) \to G(t)$ pointwisely, and $\widehat{\ell} \to \ell$, $\widehat{u} \to u$. For $(X_{n+1}, Y_{n+1}) \sim \widetilde{\mathbb{P}}$, the coverage of Algorithm 2 is $F(\widehat{v}; \widetilde{\mathbb{P}})$, where $\widehat{v} = \inf\{t\colon G(t) \ge 1 - \alpha\}$. Thus, the sharpness of our procedure relies on the difference from $G(\cdot)$ to $F(\cdot; \mathcal{P})$. The closer $G(t)$ to the worst-case c.d.f. Eq. **16**, the closer $\widehat{v}$ to $\inf\{t\colon F(t; \mathcal{P}) \ge 1 - \alpha\}$ (the worst-case quantile), hence the closer the coverage of our procedure to $1 - \alpha$. We provide exact characterizations of Eq. **16** in two problems.

C.1. Sharpness as a robust prediction problem. In the general robust prediction problem, the identification set is

$$\mathcal{P}(\mathbb{P}, \ell, u) = \left\{\widetilde{\mathbb{P}}\colon \ell(x) \le \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(x,y) \le u(x)\ \ \mathbb{P}\text{-a.s.}\right\}. \quad [17]$$

Proposition 3.2. *For each* $t \in \mathbb{R}$*, the worst-case distribution function in* Eq. **17** *is*

$$F\big(t; \mathcal{P}(\mathbb{P}, \ell, u)\big) = \max\Big\{\mathbb{E}\big[\mathbb{1}_{\{V(X,Y) \le t\}}\, \ell(X)\big],\ 1 - \mathbb{E}\big[\mathbb{1}_{\{V(X,Y) > t\}}\, u(X)\big]\Big\}, \quad [18]$$

where $\mathbb{E}$ *is the expectation over* $(X, Y) \sim \mathbb{P}$*.*

The conservative distribution function $G(\cdot)$ constructed in Eq. **13** coincides with the actual worst-case distribution function provided in Eq. **18**. Therefore, ruling out the estimation errors, the PAC-type procedure proposed in *PAC-Type Robust Conformal Inference* is indeed sharp.

C.2. Sharpness as a counterfactual inference problem. The sharpness in the counterfactual inference problem is a bit more subtle. The key difference is that the covariate shift is identifiable while only the shift in $\mathbb{P}_{Y|X}$ varies, leading to a potentially smaller identification set.

For clarity, we denote the unknown superpopulation as $(X, Y(0), Y(1), U, T) \sim \mathbb{P}^{\text{sup}}$ and the observed distribution as $(X, Y, T) \sim \mathbb{P}^{\text{obs}}$. A meaningful superpopulation $\mathbb{P}^{\text{sup}}$ should

agree with $\mathbb{P}^{\text{obs}}$ on the observable:

$$\mathbb{P}^{\text{sup}}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{X,Y,T}, \quad [19]$$

which is called the data-compatibility condition in ref. 27. Consider the counterfactual inference of $Y(1)$ for units in the control group. Letting $\mathbb{P} = \mathbb{P}_{X,Y(1)} = \mathbb{P}^{\text{obs}}_{X,Y(1)\,|\,T=1}$ be the training distribution, we have the following sharp characterization of the identification set.

Proposition 3.3. *[Identification set] For the counterfactual inference of $Y(1) \mid T = 0$ under confounding level $\Gamma \geq 1$, we define the data-compatible identification set as*

$$\mathcal{P} = \left\{\mathbb{P}^{\text{sup}}_{X,Y(1)\,|\,T=0}\colon\ \mathbb{P}^{\text{sup}} \text{ obeys [3] and [19]}\right\}.$$

Then, we have $\mathcal{P} = \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$, where

$$\mathcal{P}(\mathbb{P}, f, \ell_0, u_0) = \left\{\widetilde{\mathbb{P}}\colon\ \frac{d\widetilde{\mathbb{P}}_X}{d\mathbb{P}_X}(x) = f(x),\right.$$
$$\left.\ell_0(x) \leq \frac{d\widetilde{\mathbb{P}}_{Y(1)\,|\,X}}{d\mathbb{P}_{Y(1)\,|\,X}}(y\,|\,x) \leq u_0(x)\ \ \mathbb{P}\text{-a.s.}\right\}. \quad [20]$$

Writing $e(x) = \mathbb{P}^{\text{obs}}(T = 1 \mid X = x)$ and $\bar{p} = \mathbb{P}^{\text{obs}}(T = 1)$, we have $\ell_0(x) = 1/\Gamma$, $u_0(x) = \Gamma$, and $f(x) = \frac{\bar{p}(1-e(x))}{(1-\bar{p})\cdot e(x)}$.

Employing similar arguments, each inferential target in *Bounding the Distributional Shift* corresponds to an identification set similar to Eq. **20**, where the X-likelihood ratio is identifiable from the training distribution, and the conditional likelihood ratio can be bounded by identifiable functions ℓ_0 and u_0.

We then investigate $F(\cdot\,; \mathcal{P}(\mathbb{P}, f, \ell_0, u_0))$ with general functions f, ℓ_0, u_0 in light of Proposition 3.3. For any $x \in \mathcal{X}$ and any $\beta \in [0, 1]$, we denote the β-conditional quantile function of $\mathbb{P} = \mathbb{P}^{\text{obs}}_{X,Y(1)\,|\,T=1}$ (up to a.s. equivalence) as

$$q(\beta\,; x, \mathbb{P}) = \inf\left\{z\colon\ \mathbb{P}(Y \leq z \mid X = x) \geq \beta\right\}.$$

Proposition 3.4. *For each $t \in \mathbb{R}$, the worst-case distribution function in* Eq. **20** *is*

$$F\big(t\,; \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)\big) = \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\leq t\}}\, w^*(X, Y)\big], \quad [21]$$

where the expectation is with respect to $(X, Y) \sim \mathbb{P}$, and

$$w^*(x, y) = f(x) \cdot \big[\ell_0(x)\,\mathbb{1}_{\{V(x,y)<q(\tau(x)\,;x,\mathbb{P})\}}$$
$$+\ \gamma_0(x)\,\mathbb{1}_{\{V(x,y)=q(\tau(x)\,;x,\mathbb{P})\}} + u_0(x)\,\mathbb{1}_{\{V(x,y)>q(\tau(x)\,;x,\mathbb{P})\}}\big].$$

Here, $\tau(x) = \frac{u_0(x)-1}{u_0(x)-\ell_0(x)}$, and $\gamma_0(x)$ is chosen such that $\mathbb{E}[w^*(x, Y) \mid X = x] = f(x)$ for $\mathbb{P}$-almost all $x \in \mathcal{X}$.

As indicated by Proposition 3.4, the worst-case likelihood ratio function $w^*(x, y)$ is separated into two regions: one taking the lower bound $\ell(x)$ and the other taking the upper bound $u(x)$. This is similar to Eq. **17** with the proviso that the boundary $q(\tau(x)\,; x, \mathbb{P})$ is more complicated.

It is possible to attain sharpness (asymptotically exact $1 - \alpha$ coverage) with a few more efforts. Under the marginal Γ-selection Eq. **3**, one can check that $\tau(x) \equiv \tau$ for some constant $\tau \in (0, 1)$. Therefore, letting $G(t) = F(t\,; \mathcal{P}(\mathbb{P}, f, \ell_0, u_0))$ as in Eq. **21** yields a sharper procedure. If a tight lower bound $\widehat{G}_n(t)$ for $G(t)$ could be constructed, the worst-case coverage would equal $1-\alpha$ asymptotically. Such modifications are beyond the scope of this work. The worst-case distribution corresponds to the situation where the odds ratio Eq. **3** is either Γ or $1/\Gamma$. This is far from what may be expected in practice. *SI Appendix*, Appendix 2.D derives sharper bounds under parametric model assumptions.

D. Numerical Experiments. We apply Algorithm 2 to the same settings as of *Numerical Experiments* with a fixed confidence level $\delta = 0.05$ and the W-S–R bound as detailed in *SI Appendix*, Appendix 2.A. The empirical coverage is evaluated on 10,000 test samples for N = 1,000 independent runs. The 0.05-th quantiles of empirical coverage for various target level $1 - \alpha$ are summarized in Fig. 5 using estimated $\widehat{\ell}(\cdot)$ and $\widehat{u}(\cdot)$. Results with the ground truth $\ell(\cdot)$ and $u(\cdot)$ are in *SI Appendix*, Fig. 5 in Appendix 4.B. The marginal coverage is presented in *SI Appendix*, Figs. S3 and S4 in SI Appendix 4.B.

In Fig. 5, the 0.05-th quantiles of empirical coverage all lie above the 45°-line, which confirms the validity of our procedure. The solid lines in Fig. 5 almost overlap with the 45°-line, since by design, the data generating distribution is a near-worst case. This shows the sharpness of our method and agrees with the theoretical justification in *Sharpness*. Similar to Algorithm 1, the PAC-type algorithm is robust to the estimation error. Fig. 6, presents the gap $\widehat{\Delta}$ in Eq. **15** as well as the estimation errors. Even though the estimation (dashed lines) can be nonnegligibly off, the actual coverage gap is small.

Comparing the Two Algorithms. We take a moment here to discuss the pros and cons of the two algorithms we introduced. Algorithm 1 achieves $(1 - \alpha)$ coverage where the average is taken over the randomness of the calibration data and the test point. Algorithm 2, however, achieves valid coverage over the randomness of the test point conditional on the calibration data (with high probability). Second, a linear search suffices in Algorithm 1 (c.f. Remark 1.1) while Algorithm 2 is computationally more intensive as one needs to construct $\widehat{G}_n(t)$ at every t (which reduces to computing $\widehat{G}_n(V_i)$ for all $i \in \mathcal{D}_{\text{calib}}$). Finally, Algorithm 1 uses a uniform bound on the weight function to extract a conservative quantile. Algorithm 2 additionally takes into account that the weight function is a likelihood ratio, and this is where the sharpness comes from.

4. Sensitivity Analysis of ITEs

We are now ready to present our framework of sensitivity analysis for ITEs. We first construct prediction sets for ITEs and then invert the prediction sets to output the Γ-values. Finally, we show the statistical meaning of the Γ-values from a hypothesis testing perspective.

A. Robust Predictive Inference for ITEs. We consider two cases: 1) if only one outcome is missing, we use the prediction set for the counterfactual to form that for the ITE; 2) if both outcomes are missing, we combine prediction sets for both $Y(1)$ and $Y(0)$ to form that for the ITE.

One outcome missing. For a test sample X_{n+1} with $T_{n+1} = w$, for $w \in \{0, 1\}$, only $Y_{n+1}(w)$ is observed while $Y_{n+1}(1 - w)$ is missing. The target distribution for counterfactual prediction is $\mathbb{P}_{X,Y(1-w)\,|\,T=w}$. Using the method introduced in Algorithm 1 or 2 with bounds in *SI Appendix*, Table S1, we construct a prediction set $\widehat{C}_{1-w}(X_{n+1}, \Gamma, 1 - \alpha)$ for $Y_{n+1}(1 - w)$ with coverage level $1 - \alpha$ conditional on $T_{n+1} = w$. We then create

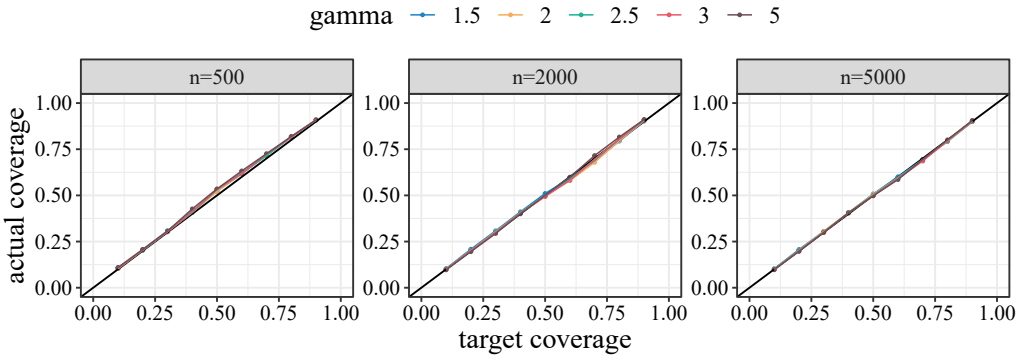

Fig. 5. 0.05-th quantile of empirical coverage in Algorithm 2 with estimated bounds. Each plot corresponds to a sample size $n = n_{\text{calib}}$ while fixing $p = 20$. Within a plot, each line represents a confounding level Γ.

the prediction set for ITE

$$\widehat{C}(X_{n+1}, \Gamma) = \begin{cases} Y_{n+1}(1) - \widehat{C}_0(X_{n+1}, \Gamma, 1-\alpha), & \text{if } w = 1, \\ \widehat{C}_1(X_{n+1}, \Gamma, 1-\alpha) - Y_{n+1}(0), & \text{if } w = 0. \end{cases}$$

Both outcomes missing. When both outcomes are missing, the target distribution is $\mathbb{P}_{X,Y(1)}$ and $\mathbb{P}_{X,Y(0)}$, and we use a Bonferroni correction to construct the predictive set for ITEs. Let X_{n+1} be a test sample with $Y_{n+1}(1)$ and $Y_{n+1}(0)$ missing. Using Algorithms 1 or 2, we construct prediction set $\widehat{C}_w(X_{n+1}, \Gamma, 1-\alpha/2, \delta/2)$, where $\delta/2$ is the confidence level for Algorithm 2. Then, we let

$$\widehat{C}(X_{n+1}, \Gamma) = \big\{y - z \colon y \in \widehat{C}_1(X_{n+1}, \Gamma, 1-\alpha/2, \delta/2) \text{ and } z \in \widehat{C}_0(X_{n+1}, \Gamma, 1-\alpha/2, \delta/2)\big\}. \quad \textbf{[22]}$$

The coverage guarantee for the above two cases directly follows from the validity of counterfactual prediction intervals; we include Propositions 2.2 and 2.3 in *SI Appendix*, Appendix 2.B for completeness.

In practice, the Bonferroni correction Eq. **22** might be too conservative. Lei and Candès (1) introduced a nested method that efficiently combines the counterfactual intervals to form the interval for the ITE; their method also applies here. We defer the details to *SI Appendix*, Appendix 2.E.

B. The Γ-Value: Inverting Nested Prediction Sets. For a new unit, we consider hypotheses indexed by $\Gamma \in [1, \infty)$:

$$H_0(\Gamma) \colon Y_{n+1}(1) - Y_{n+1}(0) \in C \text{ and } \mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma). \quad \textbf{[23]}$$

If we reject $H_0(\Gamma)$, we are saying that either $Y(1) - Y(0) \notin C$ or the confounding level of the observational data is at least Γ. The set C specifies the hypothesis, or equivalently, the causal conclusion one wishes to make. For example, setting $C = \{0\}$, we are testing whether the ITE is exactly zero; setting $C = (-\infty, 0]$, we wish to test whether the ITE is negative.

The hypothesis Eq. **23** involves the random variable $Y_{n+1}(1) - Y_{n+1}(0)$, and there are two ways to treat such hypotheses: We may either regard it as a deterministic hypothesis, which means the condition in Eq. **23** holds almost surely, or as a random hypothesis such that $H_0(\Gamma)$ is true with some probability. In both cases, the type I error is defined as $H_0(\Gamma)$ being true and rejected at the same time. Hereafter, we denote

$$\Gamma^* = \inf\big\{\Gamma \colon \mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma)\big\},$$

for the true super population $\mathbb{P}^{\text{sup}}$, and assume without loss of generality that $\mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma^*)$. Our goal is to test the set of hypotheses $\{\mathcal{H}_0(\Gamma)\}_{\Gamma \geq 1}$ simultaneously. Put

$$\mathcal{H}_0 = \{\Gamma \colon H_0(\Gamma) \text{ is true}\}.$$

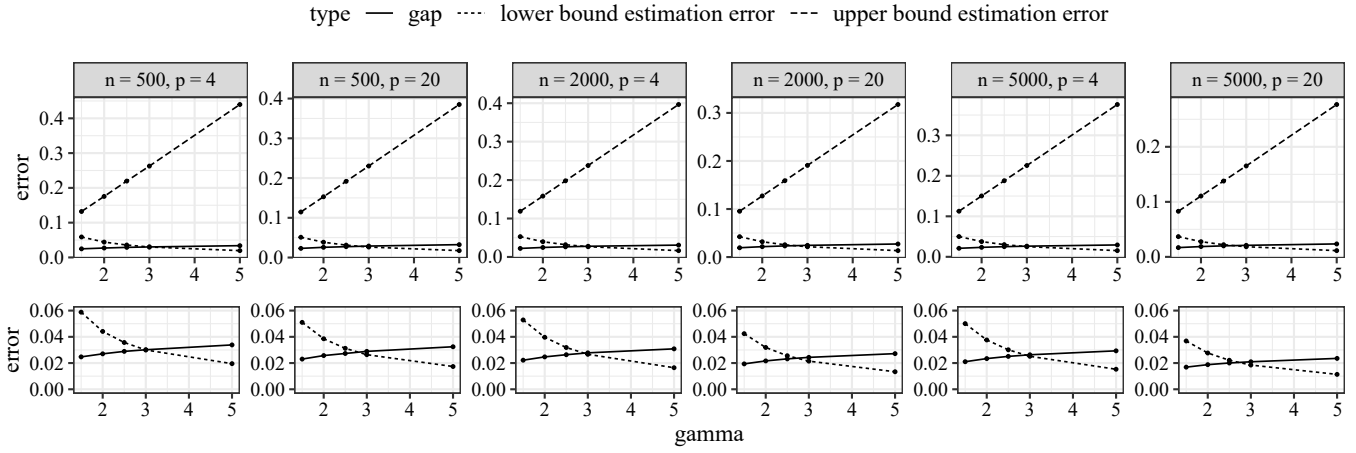

Fig. 6. Empirical gap and estimation errors. The plots in the second row zoom in on the gaps. Each plot corresponds to a sample size $n = n_{\text{calib}}$ and a dimension p. The long-dashed lines are $\|\widehat{u}(X) - u(X)\|_1$, the short-dashed lines are $\|\widehat{\ell}(X) - \ell(X)\|_1$, and the solid lines are $\widehat{\Delta}$ defined in Theorem 3.1.

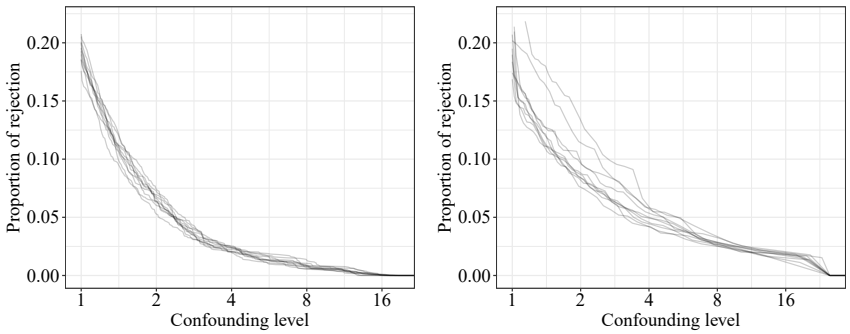

Fig. 7. Empirical survival function from testing $\{H_0^-(\Gamma)\}$ with Algorithm 1 (*Left*) and 2 (*Right*).

In the case of deterministic hypotheses, $\mathcal{H}_0$ is either an empty set (if $Y_{n+1}(1) - Y_{n+1}(0) \in C$ a.s. is false) or an interval $[\Gamma^*, \infty)$ (if $Y_{n+1}(1) - Y_{n+1}(0) \in C$ a.s. is true). In the case of random hypotheses, $\mathcal{H}_0$ is a random set—$\mathcal{H}_0$ is an empty set if $Y_{n+1}(1) - Y_{n+1}(0) \notin C$, or an interval $[\Gamma^*, \infty)$ when $Y_{n+1}(1) - Y_{n+1}(0) \in C$.

Let $\widehat{C}(X_{n+1}, \Gamma)$ be the prediction set for the ITE constructed as in *Robust Predictive Inference for ITEs* with the confounding level $\Gamma \geq 1$ and the target coverage $1 - \alpha$. In the case of one missing outcome, $\widehat{C}(X_{n+1}, \Gamma)$ implicitly depends on the observed outcome as well. Note that the prediction sets are nested in Γ in the following sense: for each fixed coverage level $\alpha \in (0, 1)$ (and confidence level δ if necessary), it holds that $\widehat{C}(X_{n+1}, \Gamma) \subset \widehat{C}(X_{n+1}, \Gamma')$ for any $\Gamma' \geq \Gamma \geq 1$. Moving on, we define the rejection set as

$$\mathcal{R} = \{\Gamma\colon\ C \cap \widehat{C}(X_{n+1}, \Gamma) = \varnothing\}.$$

That is, we reject all $H_0(\Gamma)$ for which $\widehat{C}(X_{n+1}, \Gamma)$ does not overlap with the target set C. The critical value is defined as $\widehat{\Gamma} := \sup \mathcal{R}$, with the convention that $\widehat{\Gamma} = 1$ if $\mathcal{R} = \varnothing$. $\widehat{\Gamma}$ is the formal definition of the Γ-value we introduced in Γ-*Values*. The Γ-value is a quantity specific to a unit (instead of a population quantity). Due to the variability in ITE, it might not converge to a constant as the training sample size goes to infinity.

Proposition 4.1. *[Simultaneous control] Fix a target level α (and a confidence level δ if necessary). For any $\Gamma \geq 1$, let $\widehat{C}(X_{n+1}, \Gamma)$ be the output of Algorithms 1 or 2 with the confounding level Γ. The marginal probability of making a false rejection can be controlled as*

$$\begin{aligned} mErr &:= \mathbb{P}(\mathcal{R} \cap \mathcal{H}_0 \neq \varnothing) \\ &\leq \mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}(X_{n+1}, \Gamma^*)\big), \end{aligned}$$

where the probability $\mathbb{P}$ is taken over $\mathcal{D}_{\text{calib}}$ and the new sample on both sides. Furthermore, the $\mathcal{D}_{\text{calib}}$-conditional probability of making an error satisfies

$$\begin{aligned} dErr &:= \mathbb{P}(\mathcal{R} \cap \mathcal{H}_0 \neq \varnothing \,|\, \mathcal{D}_{\text{calib}}) \\ &\leq \mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \notin \widehat{C}(X_{n+1}, \Gamma^*) \,\big|\, \mathcal{D}_{\text{calib}}\big). \end{aligned}$$

By Proposition 4.1, as long as the predictive inference achieves valid coverage at any fixed confounding level, without any adjustment of multiplicity, we achieve simultaneous control over the sequence of testing problems.

Hypothesis testing provides an interpretation of the Γ-value: the risk of $Y_{n+1}(1) - Y_{n+1}(0) \in C$ being true but rejected at $\widehat{\Gamma} \geq \Gamma^*$ is (approximately) under α. When testing deterministic hypotheses, if $Y(1) - Y(0) \in C$ almost surely, $\widehat{\Gamma}$ is a $(1 - \alpha)$ lower confidence bound for Γ^*. It is in accordance with the common practice of sensitivity analysis to find a critical value $\widehat{\Gamma}$ that inverts a causal conclusion and check whether $\widehat{\Gamma}$ is too large to be true in order to assess the robustness of such conclusion. We note that as with other sensitivity analysis tools from the literature, the Γ-value depends on a specific confidence level α.

With the general recipe of assessing robustness of causal conclusions on ITE, we provide concrete examples of the target set C and the corresponding forms of $\widehat{C}(X_{n+1}, \Gamma)$ in *SI Appendix*, Appendix 2.C. Additional simulation studies for sensitivity analysis are in *SI Appendix*, Appendix 4.C.

5. Real Data Analysis

We finally apply our procedures to an observational study dataset (50). Here, we conduct sensitivity analysis on the treated units. We consider: i) $H_0^-(\Gamma)\colon\ Y(1) - Y(0) \leq 0$ and $\Gamma^* \leq \Gamma$; rejecting the null suggests a positive ITE. ii) $H_0^+(\Gamma)\colon\ Y(1) -$

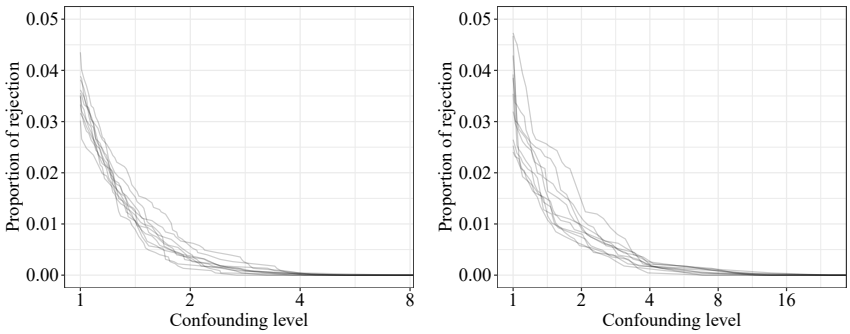

Fig. 8. Empirical survival function from testing $\{H_0^+(\Gamma)\}$ with Algorithm 1 (*Left*) and 2 (*Right*).

$Y(0) \geq 0$ and $\Gamma^* \leq \Gamma$; rejecting this null suggests a negative ITE. Our results find robust evidence for a number of positive ITEs, while few negative ITEs are detected. We also conduct counterfactual prediction with a semisynthetic dataset in *SI Appendix*, Appendix 4.D.

We randomly sample 1/3 of the original data as $\mathcal{D}_{\text{train}}$. Among the remaining, those with $T = 0$ are used as $\mathcal{D}_{\text{calib}}$, and those with $T = 1$ as the test sample $|\mathcal{D}_{\text{test}}|$. We have $|\mathcal{D}_{\text{train}}| \approx 2329$, $|\mathcal{D}_{\text{calib}}| \approx 4650$ and $|\mathcal{D}_{\text{test}}| \approx 2250$. We conduct sensitivity analysis on the dataset with $\alpha = 0.1$ and $\delta = 0.05$. The survival function for Γ-values $S(\Gamma) = \mathbb{P}(\widehat{\Gamma} > \Gamma)$ for testing $H_0^-(\Gamma)$ is evaluated in Fig. 7, which shows the evidence for positive ITEs. To illustrate the variability of the procedures, we present the empirical functions for $N = 10$ independent runs, hence the multiple curves in each plot.

Averaged over multiple independent runs, there are 19.60% or 20.46% of the treated test samples that we find at $\Gamma = 1$ as positive ITEs, using Algorithms 1 and 2, respectively. There are 6.80% or 9.65% of the test sample that we find at $\Gamma = 2$ as positive ITEs. With Algorithm 1, around 2% of the test sample have a Γ-value greater than 5. With Algorithm 2, about 2.5% of the test sample have Γ-values greater than 10, and some test samples have a Γ-value as large as 25, showing robust evidence of a positive ITE. Our framework guarantees that the mistake (i.e., rejecting an actually negative ITE at a too-large confounding level) is bounded by $\alpha = 0.1$ on average.

We then report the proportion of rejected $H_0^+(\Gamma)$ in Fig. 8, showing the evidence for negative ITEs.

Averaged over multiple runs, there are 3.58% and 3.58% of ITEs of the treated test sample that we find as negative at $\Gamma = 1$, using Algorithms 1 and 2, respectively. At $\Gamma = 2$, these proportions are 0.38% with Algorithm 1 and 1.01% with Algorithm 2. Algorithm 2 produces a little stronger evidence against unmeasured confounding, but it is slightly less stable. Very few of the samples have Γ-values larger than 2 using both algorithms.

Data, Materials, and Software Availability. Previously published data were used for this work (50). The original data can also be found in the github repository: https://github.com/ying531/cfsensitivity_paper, which also contains pre-processing codes to reproduce the analysis in the paper.

ACKNOWLEDGMENTS. The authors thank Isaac Gibbs, Kevin Guo, Suyash Gupta, Jayoon Jang, Lihua Lei, Shuangning Li, and Dominik Rothenhäusler for helpful discussions. E.J.C. was supported by the Office of Naval Research grant N00014-20-12157, the National Science Foundation grant DMS 2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594. Y.J. was partially supported by ARO grant 2003514594. Z.R. was partially supported by ONR grant N00014-20-1-2337 and NIH grants R56HG010812, R01MH113078, and R01MH123157.

Author affiliations: [a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Statistics, University of Chicago, Chicago, IL 60605; and [c]Department of Mathematics, Stanford University, Stanford, CA 94305

Author contributions: Y.J., Z.R., and E.J.C. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

1. L. Lei, E. J. Candès, Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. Ser. B* **83**, 911–938 (2021).
2. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer Science & Business Media, 2005).
3. V. Vovk, I. Nouretdinov, A. Gammerman, On-line predictive linear regression. *Ann. Stat.* **37**, 1566–1590 (2009).
4. G. W. Imbens, D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, 2015).
5. D. B. Rubin, Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* **6**, 34–58 (1978).
6. P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
7. M. Rutter, *Identifying the Environmental Causes of Disease: How Should We Decide What to Believe and When to Take Action? Report Synopsis* (Academy of Medical Sciences, 2007).
8. Z. Fewell, G. Davey Smith, J. A. Sterne, The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *Am. J. Epidemiol.* **166**, 646–655 (2007).
9. M. Gaudino *et al.*, Unmeasured confounders in observational studies comparing bilateral versus single internal thoracic artery for coronary artery bypass grafting: A meta-analysis. *J. Am. Heart Assoc.* **7**, e008010 (2018).
10. J. Neyman, Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10**, 1–51 (1923).
11. S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, L. Tian, Bounds on the conditional and average treatment effect with unobserved confounding factors. *Ann. Statist.* **50**, 2587–2615 (2022).
12. D. R. Cox, *Planning of Experiments* (Wiley, 1958).
13. D. B. Rubin, Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat. Sci.* **5**, 472–480 (1990).
14. Q. Zhao, D. S. Small, B. B. Bhattacharya, Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Ser. B* **81**, 735–761 (2019).
15. Z. Tan, A distributional approach for causal inference using propensity scores. *J. Am. Stat. Assoc.* **101**, 1619–1637 (2006).
16. J. Cornfield *et al.*, Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22**, 173–203 (1959).
17. P. R. Rosenbaum, Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74**, 13–26 (1987).
18. I. D. Bross, Spurious effects from an extraneous variable. *J. Chronic Dis.* **19**, 637–647 (1966).
19. I. D. Bross, Pertinency of an extraneous variable. *J. Chronic Dis.* **20**, 487–495 (1967).
20. P. R. Rosenbaum, D. B. Rubin, Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc.: Ser. B (Methodol.)* **45**, 212–218 (1983).
21. J. L. Gastwirth, A. M. Krieger, P. R. Rosenbaum, Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85**, 907–920 (1998).
22. P. R. Rosenbaum, Attributing effects to treatment in matched observational studies. *J. Am. Stat. Assoc.* **97**, 183–192 (2002).
23. P. Rosenbaum, *Observational Studies. Springer Series in Statistics* (Springer, 2002).
24. G. W. Imbens, Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93**, 126–132 (2003).
25. P. Ding, T. J. VanderWeele, Sensitivity analysis without assumptions. Epidemiol. (Cambridge Mass.) **27**, 368 (2016).
26. T. J. VanderWeele, P. Ding, Sensitivity analysis in observational research: Introducing the E-value. *Ann. Int. Med.* **167**, 268–274 (2017).
27. J. Dorn, K. Guo, Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *J. Am. Stat. Assoc.* (2022). DOI: 10.1080/01621459.2022.2069572.
28. H. Namkoong, R. Keramati, S. Yadlowsky, E. Brunskill, "Off-policy policy evaluation for sequential decisions under unobserved confounding" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 18819–18831. https://proceedings.neurips.cc/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf.
29. N. Kallus, A. Zhou, Minimax-optimal policy learning under unobserved confounding. *Manag. Sci.* **67**, 2870–2890 (2021).
30. A. Gammerman, V. Vovk, Hedging predictions in machine learning. *Comput. J.* **50**, 151–163 (2007).
31. G. Shafer, V. Vovk, A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**, 371–421 (2008).
32. V. Vovk, "Conditional validity of inductive conformal predictors in Asian conference on machine learning" in *PMLR* (2012), pp. 475–490.
33. V. Vovk, "Transductive conformal predictors" in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2013), pp. 348–360.
34. W. Hoeffding, The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23**, 169–192 (1952).
35. J. Lei, L. Wasserman, Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc.: Ser. B: Stat. Methodol.* **76**, 71–96 (2014).
36. J. Lei, M. GSell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
37. J. Lei, Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika* **106**, 749–764 (2019).
38. Y. Romano, M. Sesia, E. J. Candès, "Classification with valid and adaptive coverage" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., , 2020), vol. 33, pp. 3581–3591. https://proceedings.neurips.cc/paper/2020/file/244edd7e85dc81602b7615cd705545f5-Paper.pdf.
39. M. Cauchois, S. Gupta, J. C. Duchi, Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.* **22**, 1–42 (2021).
40. M. Cauchois, S. Gupta, A. Ali, J. C. Duchi, Robust validation: Confident predictions even when distributions shift. arXiv [Preprint] (2020). http://arxiv.org/abs/2008.04267arXiv:2008.04267 (Accessed 10 August 2020).
41. S. Park, E. Dobriban, I. Lee, O. Bastani, Pac prediction sets under covariate shift. arXiv [Preprint] (2021). http://arxiv.org/abs/2106.09848 (Accessed 17 June 2021).
42. M. Yin, C. Shi, Y. Wang, D. M. Blei, Conformal sensitivity analysis for individual treatment effects (2021).

43. R. J. Tibshirani, R. F. Barber, E. J. Candès, A. Ramdas, "Conformal prediction under covariate shift" in *Advances in Neural Information Processing Systems* (2019), vol. 32.
44. C. Gupta, A. K. Kuchibhotla, A. K. Ramdas, Nested conformal prediction and quantile out-of-bag ensemble methods. arXiv [Preprint] (2019). http://arxiv.org/abs/1910.10562 (Accessed 23 October 2019).
45. Y. Romano, E. Patterson, E. Candes, Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**, 3543–3553 (2019).
46. N. Meinshausen, G. Ridgeway, Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999 (2006).
47. S. Bates, A. Angelopoulos, L. Lei, J. Malik, M. Jordan, Distribution-free, risk-controlling prediction sets. *J. ACM* **68**, Article 43 (2021) p. 34, https://doi.org/10.1145/3478535.
48. S. Bates, E. Candès, L. Lei, Y. Romano, M. Sesia, Testing for outliers with conformal P-values. arXiv [Preprint] (2021). http://arxiv.org/abs/2104.08279 (Accessed 16 April 2021).
49. I. Waudby-Smith, A. Ramdas, Estimating means of bounded random variables by betting (2021).
50. C. Carvalho, A. Feller, J. Murray, S. Woody, D. Yeager, Assessing treatment effect variation in observational studies: Results from a data challenge. *Observ. Stud.* **5**, 21–35 (2019).