Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Conformal Sensitivity Analysis for Individual Treatment Effects

Mingzhang Yin, Claudia Shi, Yixin Wang & David M. Blei

To cite this article: Mingzhang Yin, Claudia Shi, Yixin Wang & David M. Blei (2022): Conformal Sensitivity Analysis for Individual Treatment Effects, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2102503

To link to this article: https://doi.org/10.1080/01621459.2022.2102503

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2022.2102503

Taylor & Francis
Taylor & Francis Group

Conformal Sensitivity Analysis for Individual Treatment Effects

Mingzhang Yin[a], Claudia Shi[b], Yixin Wang[c], and David M. Blei[d]

[a]Warrington College of Business, University of Florida, Gainesville, FL; [b]Department of Computer Science, Columbia University, New York, NY; [c]Department of Statistics, University of Michigan, Ann Arbor, MI; [d]Department of Computer Science and Department of Statistics, Columbia University, New York, NY

ABSTRACT

Estimating an individual treatment effect (ITE) is essential to personalized decision making. However, existing methods for estimating the ITE often rely on unconfoundedness, an assumption that is fundamentally untestable with observed data. To assess the robustness of individual-level causal conclusion with unconfoundedness, this article proposes a method for sensitivity analysis of the ITE, a way to estimate a range of the ITE under unobserved confounding. The method we develop quantifies unmeasured confounding through a marginal sensitivity model, and adapts the framework of conformal inference to estimate an ITE interval at a given confounding strength. In particular, we formulate this sensitivity analysis as a conformal inference problem under distribution shift, and we extend existing methods of covariate-shifted conformal inference to this more general setting. The resulting predictive interval has guaranteed nominal coverage of the ITE and provides this coverage with distribution-free and nonasymptotic guarantees. We evaluate the method on synthetic data and illustrate its application in an observational study. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2021
Accepted July 2022

KEYWORDS
Distribution shift; Predictive inference; Uncertainty quantification; Unconfoundedness

1. Introduction

Consider a person who ponders whether to take the COVID-19 vaccine. She is interested in understanding how much risk the COVID vaccine can reduce for her. However, most large-scale observational studies are conducted to estimate the average vaccine efficacy over a whole population (Haas et al. 2021), and such population-level estimates provide a summary that cannot reflect individual heterogeneity.

The causal estimand that captures individual heterogeneity is the individual treatment effect (ITE), the per-individual difference between the potential outcomes. However, estimation of the ITE is fundamentally challenging, even beyond the usual population-level estimands, because of its inherent uncertainty. To address this challenge, researchers have recently adapted the method of conformal inference (Vovk, Gammerman, and Shafer 2005) to estimate ITE intervals with good theoretical guarantees (Kivaranovic et al. 2020b; Lei and Candès 2021). Conformal inference helps estimate an interval that contains the true ITE with a guaranteed minimal probability.

Conformal inference for ITE estimation is an important innovation, but it comes with assumptions. In particular, it relies on the usual assumption of unconfoundedness (Kivaranovic et al. 2020b; Lei and Candès 2021) that the treatment assignment is conditionally independent of the potential outcomes. In practice, this assumption can be difficult to accept for many observational studies (Greenland, Pearl, and Robins 1999), and violations of unconfoundedness will introduce hidden biases into the estimation of the ITE. In the context of COVID-19 vaccine studies, for example, unmeasured confounding may come from coexisting conditions, medical resources, and socioeconomic status (Amin-Chowdhury and Ladhani 2021).

To assess the robustness of individual-level causal conclusions with unconfoundedness, this article develops a method for sensitivity analysis of the ITE. The idea of sensitivity analysis is to quantify the violation of the required assumptions and then to produce intervals of causal estimates that account for such violations. In the context of the ITE, a sensitivity analysis must account for two sources of uncertainty: the inherent uncertainty of the estimand itself and the uncertainty due to violations of the required assumptions.

We develop conformal sensitivity analysis (CSA), a method for sensitivity analysis of ITE interval estimation. Given a prespecified amount of unmeasured confounding, CSA estimates an interval that captures the true ITE with a guaranteed probability. We develop CSA by relaxing the assumption of unconfoundedness with a marginal sensitivity model (MSM) (Rosenbaum 2002; Tan 2006), a general model of the treatment assignment and potential outcomes that includes a real-valued parameter for the strength of unmeasured confounding. With the MSM in hand, we then show how sensitivity analysis can be formulated as a predictive inference of the missing potential outcomes, but under a general distribution shift. Finally, we extend weighted conformal prediction (Tibshirani et al. 2019), a predictive inference method developed in the setting of covariate shift, to this more general setting of distribution shift.

CONTACT Mingzhang Yin *mzyin11@gmail.com* Warrington College of Business, Marketing Department, University of Florida, Gainesville, FL. Initial work is done at Data Science Institute and Irving Institute for Cancer Dynamics, Columbia University, New York, NY.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

The CSA algorithm contains two stages. Given a specification of an MSM, it first computes the range of weight functions of the weighted conformal prediction. While covariate shift leads to a single weight function, distribution shift requires a range. Then it uses these functions to quantify the bounds of the ITE, found by solving an optimization problem with constrained weights. The resulting algorithm provides a valid interval estimate of an ITE whenever the true data generation is consistent with the MSM.

CSA has several practical and theoretical strengths. By leveraging conformal inference, it makes minimal assumptions about the underlying distribution of the observed data, and its theoretical guarantees are valid with finite data. By using an MSM, it does not impose additional untestable assumptions over the distribution of a latent confounder and its effects on other variables. CSA can be used with any predictive functions to fit the treatment and outcome, and it can be applied after fitting such functions with a light computational cost.

1.1. Related Work

Conformal Inference. The framework of conformal inference was pioneered by Vladimir Vovk and his collaborators (Papadopoulos et al. 2002; Vovk, Gammerman, and Shafer 2005; Vovk, Nouretdinov, and Gammerman 2009; Vovk 2012). Recent developments of conformal inference improve its accuracy (Lei et al. 2018; Romano, Patterson, and Candès 2019), efficiency (Lei, Rinaldo, and Wasserman 2015), and extend its applicable domains (Lei, Rinaldo, and Wasserman 2015; Candès, Lei, and Ren 2021).

First, regarding accuracy, a variety of conformal inference algorithms were proposed to reduce the length of predictive band. Some algorithms rely on the conditional quantile regression of the outcome given the covariates to capture the individual heterogeneity (Romano, Patterson, and Candès 2019; Kivaranovic, Johnson, and Leeb 2020a; Sesia and Candès 2020), some adapt to skewed data by estimating the conditional histograms (Sesia and Romano 2021), and others estimate the conditional density function to produce nonconvex predictive bands (Izbicki, Shimizu, and Stern 2020; Hoff 2021). Second, to improve efficiency, the split conformal inference framework is proposed; it uses data splitting to avoid multiple refitting of the predictor (Papadopoulos 2008; Shafer and Vovk 2008; Lei and Wasserman 2014). Such data-splitting will also be adopted in this article. Finally, regarding domain extensions, the weighted conformal prediction is proposed to handle non-iid data (Tibshirani et al. 2019), generalizing conformal inference from exchangeable data to data with covariates shift.

Sensitivity analysis. Sensitivity analysis dates back to the study of the average treatment effect (ATE) of smoking on lung cancer (Cornfield et al. 1959). More recent advances for sensitivity analysis posit a hypothetical latent confounder and evaluate its impact on a causal conclusion (Rosenbaum and Rubin 1983a; Imbens 2003; Ding and VanderWeele 2016; Dorie et al. 2016; Cinelli and Hazlett 2020; Veitch and Zaveri 2020; Hong, Yang, and Qin 2021). Though intuitive, introducing a latent confounder often entails additional untestable assumptions (Franks, D'Amour, and Feller 2019). As an alternative, some methods directly model the dependency between treatment assignment and potential outcomes given the covariates, such as the MSM in this article (Robins, Rotnitzky, and Scharfstein 2000; Tan 2006). With this strategy, some sensitivity models focus on modeling the potential outcome given the treatment (Brumback et al. 2004; Blackwell 2014), while others focus on modeling the treatment distribution given the potential outcomes (Tan 2006; Yadlowsky et al. 2018; Franks, D'Amour, and Feller 2019; Zhao, Small, and Bhattacharya 2019).

Other papers consider different frameworks to evaluate the sensitivity of a causal estimate. Some sensitivity analysis methods measure the association between a latent confounder and the treatment or outcome that produces a specific amount of estimation bias (Imbens 2003; Ding and VanderWeele 2016; Cinelli and Hazlett 2020; Veitch and Zaveri 2020). Another way to evaluate sensitivity is to compute an interval estimate of the target estimand for a specific level of unmeasured confounding. For the ATE, the percentile bootstrap produces a partial identified region with asymptotically valid coverage (Zhao, Small, and Bhattacharya 2019). For the conditional average treatment effect (CATE), data-dependent interval estimations have been proposed via nonparametric and (semi-)parametric approaches (Yadlowsky et al. 2018; Kallus, Mao, and Zhou 2019; Jesson et al. 2021). This work further explores the interval estimation of the ITE under an unmeasured confounding.

In an independent and concurrent paper, Jin, Ren, and Candès (2021) also develops sensitivity analysis procedures for the ITE based on robust conformal inference. Jin, Ren, and Candès (2021) derives a sensitivity analysis based on the MSM and proposes an extended conformal inference algorithm that is equivalent to Algorithm 1. However, this article and Jin, Ren, and Candès (2021) propose the methods of analysis that are complementary and offer different perspectives. The present article defines the MSM without explicitly having to posit a latent confounder and uses Tukey's factorization for an alternative derivation of Lemma 1 (same as Lem. 3.1 Jin, Ren, and Candès 2021). We propose and implement an algorithm to improve the sharpness of the predictive set, provide tools of calibration from observed data, and design methods to evaluate the estimation over different sensitivity models in the MSM.

2. Conformal Inference of Individual Treatment Effects

We first set up the problem of ITE estimation. Next, we formulate the ITE estimation in observational study as a conformal inference problem under distribution shift, and introduce existing estimation methods under the assumption of unconfoundedness. Then we discuss the challenges presented to ITE estimation when there is unmeasured confounding.

2.1. Problem Setup

Consider N statistical units. Each unit $i \in \{1, 2, \ldots, N\}$ is associated with a tuple of random variables $(X_i, T_i, Y_i(0), Y_i(1))$.

Algorithm 1: CSA for Estimating an Unobserved Potential Outcome

Input: Data $\mathcal{Z} = (X_i, Y_i, T_i)_{i=1}^N$, where Y_i is missing if $T_i = 1 - t$; level α, sensitivity parameter Γ, target point covariates X.

Step I: Preliminary processing

1: Split the data into 2-fold $\mathcal{Z}_{\text{pre}}$ and $\mathcal{Z}_{\text{cal}}$; let $\mathcal{I}_{\text{pre}} = \{i : Z_i \in \mathcal{Z}_{\text{pre}}, T_i = t\}$, $\mathcal{I}_{\text{cal}} = \{i : Z_i \in \mathcal{Z}_{\text{cal}}, T_i = t\}$
2: Estimate propensity score $\widehat{e}(x)$ on $\mathcal{Z}_{\text{pre}}$
3: Estimate predictor $\widehat{\mu}(\cdot)$ on $\{X_i, Y_i\}_{i\in\mathcal{I}_{\text{pre}}}$

Step II: Predictive interval for $Y_i(t)$ at the target point

1: Compute nonconformity scores $\mathcal{V} = \{V_i\}_{i\in\widetilde{\mathcal{I}}_{\text{cal}}}$, $\widetilde{\mathcal{I}}_{\text{cal}} = \mathcal{I}_{\text{cal}} \cup \{N+1\}$, $V_{N+1} = \infty$
2: Compute the bounds $(w_{lo}^{\Gamma}(X_i), w_{hi}^{\Gamma}(X_i))$ for $i \in \widetilde{\mathcal{I}}_{\text{cal}}$ by Equation (18), $X_{N+1} = X$;
3: For $i \in \widetilde{\mathcal{I}}_{\text{cal}}$, initialize the weights $w_i = w_{lo}^{\Gamma}(X_i)$
4: Sort $\mathcal{V}$ in ascending order and relabel the ordered elements from 1 to $|\mathcal{V}|$
5: Relabel $\{w_i\}_{i\in\widetilde{\mathcal{I}}_{\text{cal}}}$, $\{X_i\}_{i\in\widetilde{\mathcal{I}}_{\text{cal}}}$ according to the labels of the sorted $\mathcal{V}$; set $k = |\mathcal{V}|$
6: **do**
$w_k \leftarrow w_{hi}^{\Gamma}(X_k)$
Compute normalized weights $p_i = \frac{w_i}{\sum_{j=1}^{|\mathcal{V}|} w_j}$, for $i \in \{1, 2, \ldots, |\mathcal{V}|\}$
$k \leftarrow k - 1$
while $\sum_{i=k+1}^{|\mathcal{V}|} p_i < \alpha$

Output: Compute $\widehat{C}_t^{\Gamma}(X)$ by Equation (20) with $\widehat{Q}(Z_{1:n}, X) = V_{k+1}$

$X_i \in \mathcal{X} \subset \mathbb{R}^p$ is a vector of covariates, $T_i \in \{0, 1\}$ is the treatment, $Y_i(1), Y_i(0) \in \mathcal{Y} \subset \mathbb{R}$ are the potential outcomes under treatment and control (Neyman 1923; Rubin 1974). We use $\mathbb{P}_0(X, T, Y(0), Y(1))$ to denote the true joint distribution of these variables.

We make the stable unit treatment value assumption (SUTVA) (Rubin 1980). Under SUTVA, the observed outcome $Y_i \in \mathbb{R}$ is one of the potential outcomes $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

Assumption 1 (SUTVA). There is no interference between units, and there are no unrepresented treatments.

We further assume that each unit has a positive probability of being assigned to all treatment groups and the probability is bounded away from the extremes (Rosenbaum and Rubin 1983b).

Assumption 2 (Strong overlap). $\exists\, \eta > 0, \eta < p(T_i = 1 \mid X_i) < 1 - \eta$ with probability 1.

The causal estimand of interest is the ITE. The ITE of unit i is defined as the difference between its potential outcomes, $\tau_i = Y_i(1) - Y_i(0)$.

Estimating the ITE is challenging. The fundamental problem of causal inference is that we can at most observe one potential outcome of each unit (Holland 1986). Therefore, the ITE, which requires knowing both the potential outcomes, can never be observed. Furthermore, unlike population-level causal estimands, an ITE is inherently random. Even with a known joint distribution $\mathbb{P}_0(X, T, Y(0), Y(1))$, an ITE is not point-identifiable (Hernan and Robins 2010).

To tackle these challenges, the problem of ITE estimation has been reframed as a predictive inference problem (Kivaranovic et al. 2020b; Lei and Candès 2021).

2.2. Predictive Inference in Observational Studies

The idea of predictive inference is to form a covariate-dependent predictive band that contains the outcome of a new data point with a guaranteed probability (Vovk, Gammerman, and Shafer 2005; Barber et al. 2021b). For predicting potential outcomes, predictive inference aims to use observed data $\{X_i, Y_i(t)\}_{i:T_i=t}$ from the treatment group t to learn a mapping from the covariates X to an interval estimate $\widehat{C}_t(X) \subset \mathbb{R}$ of the potential outcome $Y(t)$. For a new data point $(X, Y(t)) \sim \mathbb{P}(X, Y(t))$, the band must have a valid coverage probability $\mathbb{P}(Y(t) \in \widehat{C}_t(X)) \geq 1 - \alpha$, where the probability is taken over both X and $Y(t)$ and $\alpha \in [0, 1]$ is a predetermined level.

Conformal inference is a collection of methods that realize the goal of predictive inference (Vovk, Gammerman, and Shafer 2005). Classic conformal inference assumes the training data and the target data are *exchangeable*. Based on exchangeability, the predictive band is constructed by the quantiles of the prediction residuals. Weighted conformal prediction (WCP) extends the setting to covariate shift (Tibshirani et al. 2019), where distribution of the covariates $\mathbb{P}(X)$ changes from the training data to the target data but the outcome distribution $\mathbb{P}(Y(t) \mid X)$ remains the same. Under covariate shift, WCP produces a valid predictive interval (Tibshirani et al. 2019).

WCP has been applied to the ITE estimation (Lei and Candès 2021). Suppose we want to estimate the missing outcome $Y(t)$ of a randomly sampled unit. The relationship between the observed data and the inference target $(X, Y(t))$ is

$$\begin{aligned} &\text{Training: } (X_i, Y_i(t)) \overset{\text{iid}}{\sim} p(X \mid T = t) \cdot p(Y(t) \mid X, T = t), \\ &\qquad\qquad\qquad\qquad i \in \{i : T_i = t\}; \\ &\text{Target: } (X, Y(t)) \sim p(X) \cdot p(Y(t) \mid X). \end{aligned} \tag{1}$$

For the training data, we observe both the covariates X_i and outcome $Y_i(t)$. For a target data point, we only observe covariates X and the goal is to infer the missing outcome $Y(t)$.

WCP for the ITE estimation crucially relies on the assumption of *unconfoundedness* (Kivaranovic et al. 2020b; Lei and Candès 2021), that the units are assigned to the treatment groups based only on the observed covariates, that is, $(Y_i(0), Y_i(1)) T_i \mid X_i$. Under unconfoundedness, the conditional distributions of a potential outcome remain invariant across treatment groups, $P(Y(t) \mid X, T = t) = P(Y(t) \mid X)$. And Equation (1) becomes

$$\begin{aligned} &\text{Training: } (X_i, Y_i(t)) \overset{\text{iid}}{\sim} p(X \mid T = t) \cdot p(Y(t) \mid X), \\ &\qquad\qquad\qquad\qquad i \in \{i : T_i = t\}; \end{aligned}$$

$$\text{Target:}\quad (X, Y(t)) \sim p(X) \cdot p(Y(t) \mid X). \tag{2}$$

The only difference between the training and the target distribution is on the covariates distribution. In other words, unconfoundedness reduces the setting of counterfactual inference from the general distribution shift in Equation (1) to covariate shift in Equation (2), for which WCP is readily applicable.

2.3. ITE Estimation under Unconfoundedness

We explain how WCP tackles the predictive inference problem in Equation (1) and point out the challenges when unconfoundedness is violated.

Denote each training data pair as $Z_i := (X_i, Y_i(t))$, $Z_{1:n} = (Z_1, \ldots, Z_n)$. Given a predictive function $\widehat{\mu}(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$, conformal inference uses a scalar-valued function $V : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to measure the predictive error. For instance, V_i might be chosen as the absolute residual function $V(X_i, Y_i(t)) = |Y_i(t) - \widehat{\mu}(X_i)|$ with mean prediction $\widehat{\mu}(\cdot)$ (Vovk, Gammerman, and Shafer 2005); we will discuss how to fit the predictor $\widehat{\mu}(\cdot)$ later and take it as a fixed mapping for now. The *nonconformity score* is defined for each data point as $V_i := V(Z_i)$.

Denote the αth quantile of a random variable $X \sim p(X)$ as $Q_\alpha(X)$, where $Q_\alpha(X) = \inf\{x : p(X \le x) \ge \alpha\}$, $\alpha \in [0, 1]$. Let $\delta_v(V)$ be the Dirac delta function, defined as $\delta_v(V) = 1$ if $V = v$ and $\delta_v(V) = 0$ otherwise. We denote the quantile of a discrete distribution and the quantile of an empirical distribution as

$$Q_\alpha\left(\sum_{i=1}^{n} p_i \delta_{v_i}\right) := Q_\alpha(V),\ V \sim \sum_{i=1}^{n} p_i \delta_{v_i};$$
$$Q_\alpha(v_{1:n}) := Q_\alpha\left(\sum_{i=1}^{n} \frac{1}{n} \delta_{v_i}\right).$$

We define the *conformal weights* as the density ratio of the training and target distributions,

$$w_t(x, y) := \frac{p(X = x)p(Y(t) = y \mid X = x)}{p(X = x \mid T = t)p(Y(t) = y \mid X = x, T = t)}. \tag{3}$$

Then for units $1 \le i \le n$, let the normalized weights be

$$p_i^t(Z_{1:n}, (x, y)) := \frac{w_t(Z_i)}{\sum_{i=1}^{n} w_t(Z_i) + w_t(x, y)},$$
$$p_{n+1}^t(Z_{1:n}, (x, y)) := \frac{w_t(x, y)}{\sum_{i=1}^{n} w_t(Z_i) + w_t(x, y)}.$$

where p_i^t are the weights for the observed data and p_{n+1}^t is the weight for a new data. When the conformal weights in Equation (3) are known or computable, we can use the WCP to derive a predictive interval (Tibshirani et al. 2019),

$$\widehat{C}_t(x) = \Big\{y \in \mathbb{R} : V(x, y) \le Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i(Z_{1:n}, (x, y))\delta_{V_i} + p_{n+1}(Z_{1:n}, (x, y))\delta_\infty\Big)\Big\}. \tag{4}$$

The interval $\widehat{C}_t(x)$ is guaranteed with a preset $1 - \alpha$ coverage probability (Tibshirani et al. 2019; Lei and Candès 2021), that is, $\mathbb{P}_{(X,Y(t)) \sim p(X)p(Y(t) \mid X)}(Y(t) \in \widehat{C}_t(X)) \ge 1 - \alpha$.

Computing the predictive interval $\widehat{C}_t(x)$ relies on the conformal weights being accessible. In an ideal randomized controlled trial (RCT) with perfect compliance, the training and target data in Equation (1) are iid, hence, the conformal weights $w_t(x, y) \equiv 1$. In an observational study under unconfoundedness, the conformal weights $w_t(x, y) = p(X = x)/p(X = x|T = t)$ can be estimated from the observed data (Lei and Candès 2021).

However, when unconfoundedness is violated, the joint distribution of the covariates and outcome shifts from training to target as shown in Equation (1). We will see that the conformal weights $w_t(x, y)$ are nonidentifiable under such distribution shift. When unconfoundedness is violated, existing conformal inference cannot be directly applied to ITE estimation.

3. Sensitivity Analysis for ITEs

In this section, we develop an individual-level sensitivity analysis for estimating a missing outcome and generalize it to a sensitivity analysis for the ITE. We first define what it means to deviate from unconfoundedness. We then show how to incorporate the uncertainty from an unknown confounding into the construction of a valid predictive interval.

3.1. Confounding Strength and the Marginal Sensitivity Model

A sensitivity analysis quantifies the deviation from unconfoundeness and evaluates the corresponding range of causal estimates. We quantify the strength of unmeasured confounding by the marginal sensitivity model (MSM) (Tan 2006; Zhao, Small, and Bhattacharya 2019). Under unconfoundeness, the *propensity score*, $e(x) := p(T = 1 \mid X = x)$ (Rosenbaum and Rubin 1983b), is the same as the *selection score*, $s_t(x, y) := p(T = 1 \mid X = x, Y(t) = y)$ (Scharfstein, Rotnitzky, and Robins 1999; Robins, Rotnitzky, and Scharfstein 2000). Without unconfoundeness, the selection scores no longer equal to the propensity score. Their difference represents the strength of confounding, which can be measured by the odds ratio (OR), $\mathrm{OR}(s_t(x, y), e(x)) := \left[s_t(x, y)/(1 - s_t(x, y))\right]\left[e(x)/(1 - e(x))\right]$. The MSM assumes that under the true data distribution $\mathbb{P}_0$, the odds ratio between the selection score and the propensity score is bounded by a given range (Tan 2006; Zhao, Small, and Bhattacharya 2019).

Definition 1 (Marginal Sensitivity Model). Under the distribution $\mathbb{P}_0$ over $(X, T, Y(1), Y(0))$, assume the propensity score $e(x)$ and selection score $s_t(x, y)$ satisfy $s_t(x, y) \in \mathcal{E}(\Gamma)$, where

$$\mathcal{E}(\Gamma) = \{s(x, y) : s(x, y) : \mathcal{X} \times \mathcal{Y} \mapsto [0, 1], \\ 1/\Gamma \le \mathrm{OR}(s(x, y), e(x)) \le \Gamma, \\ \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}\}, \tag{5}$$

and the sensitivity parameter $\Gamma \ge 1$.

The magnitude of Γ is the degree of deviation from unconfoundedness. The set $\mathcal{E}(\Gamma)$ expands with an increasing Γ, representing more possible ways of the treatment assignment that are not explained by the observed covariates. By specifying Γ, the

MSM assumes a rich set of data generating distributions, which avoids imposing parametric assumptions on how an unobserved confounder interacts with the treatment and outcome.

Under the assumption of an MSM, we will develop CSA in a two-stage approach. In the first stage, we quantify how the uncertainty from unmeasured confounding propagates to the conformal weights. In the second stage, we leverage the uncertainty in the conformal weights to create a valid predictive interval.

3.2. From Confounding Strength to the Conformal Weights

We illustrate that under unmeasured confounding, the predictive interval in Equation (4) cannot be computed by the observational data, which is the main challenge in applying conformal inference for sensitivity analysis.

The conformal weights in Equation (3) decompose to two terms. The first term $p(X)/p(X \mid T = t) = p(T = t)/p(T = t \mid X)$ can be inferred from data. The second term is,

$$\frac{p(Y(t) \mid X)}{p(Y(t) \mid X, T = t)} = p(T = t \mid X) + \frac{p(Y(t) \mid X, T = 1 - t)}{p(Y(t) \mid X, T = t)} p(T = 1 - t \mid X). \tag{6}$$

Without unconfoundedness, the density ratio on the right hand side of Equation (6) involves the nonidentifiable distribution $p(Y(t) \mid X, T = 1 - t)$ of the missing potential outcome.

To deal with this challenge, we transfer the uncertainty from the unknown confounding to the uncertainty of the conformal weights. The nonidentifiable density ratio term in the conformal weights is related to the odds ratio in the MSM. Applying Bayes's rule,

$$\begin{aligned} p(T = 1 \mid X, Y(t)) =& \frac{p(Y(t) \mid X, T = 1)p(T = 1 \mid X)}{p(Y(t) \mid X)} \\ =& 1/\Big(1 + \frac{1 - e(X)}{e(X)} \frac{p(Y(t) \mid X, T = 0)}{p(Y(t) \mid X, T = 1)}\Big). \end{aligned} \tag{7}$$

Equation (7) is also known as Tukey's factorization (Brook 1964; Franks, Airoldi, and Rubin 2016; Franks, D'Amour, and Feller 2019).

Based on Tukey's factorization, the following lemma shows that the conformal weight is proportional to the inverse selection score and the density ratio of a potential outcome in the two treatment groups is bounded by the sensitivity parameter.

Lemma 1. (i) For the conformal weights in Equation (3), we have $w_t(x, y) = p(T = t)/p(T = t \mid X = x, Y(t) = y)$. (ii) The MSM with parameter Γ equivalently assumes

$$\frac{1}{\Gamma} \leq \frac{p(Y(t) = y | X = x, T = 1)}{p(Y(t) = y | X = x, T = 0)} = \mathrm{OR}(s(x, y), e(x)) \leq \Gamma \tag{8}$$

The sensitivity parameter of the MSM specifies the range of plausible conformal weights.

Lemma 2. Given an MSM with sensitivity parameter $1 \leq \Gamma < \infty$, the weight function for the weighted conformal prediction in Equation (3) is bounded by

$$\begin{aligned} &\Big(1 + \frac{1}{\Gamma}\big(\frac{1 - e(x)}{e(x)}\big)^{2t-1}\Big)p(T = t) \\ &\quad \leq w_t(x, y) \leq \Big(1 + \Gamma\big(\frac{(1 - e(x))}{e(x)}\big)^{2t-1}\Big)p(T = t). \end{aligned} \tag{9}$$

Note that the bounds in Equation (9) are uniform for all y. When $\Gamma = 1$, the upper and lower bounds of the conformal weights are the same. When $\Gamma > 1$, the conformal weights cannot be point identified. The range in Equation (9) represents the weight uncertainty.

3.3. From Conformal Weights to the Predictive Band

We first define a valid predictive band in sensitivity analysis. Then we demonstrate the validity of a predictive interval given a specific sensitivity model. The union of such intervals becomes a valid predictive band for the MSM. Finally, as a practical way to obtain the union set, we propose and solve a constrained quantile optimization problem.

Valid Predictive Bands under Sensitivity Models. By Equation (8), each selection score $s_t(x, y)$ specifies a missing outcome distribution

$$\begin{aligned} &p^{(s_t)}(Y(t) \mid X, T = 1 - t) \\ &\quad = \mathrm{OR}(s_t(X, Y(t)), e(X))^{1-2t} \cdot p(Y(t) \mid X, T = t). \end{aligned} \tag{10}$$

For a selection score s_t in the collection of sensitivity models $\mathcal{E}(\Gamma)$, the target data in Equation (1) is generated by a distribution depending on s_t, that is,

$$\begin{aligned} p^{(s_t)}(Y(t) \mid X) = &\, p(T = t \mid X)p(Y(t) \mid X, T = t) \\ &+ p(T = 1 - t \mid X)p^{(s_t)}(Y(t) \mid X, T = 1 - t). \end{aligned} \tag{11}$$

With the notation above, the validity of the predictive interval is defined as a worst-case guarantee under *all* plausible sensitivity models in $\mathcal{E}(\Gamma)$.

Definition 2. Under a set of sensitivity models $\mathcal{E}(\Gamma)$, the predictive band for the potential outcome $Y(t)$ with $(1 - \alpha)$ marginal coverage is a band that satisfies

$$\mathbb{P}_{X, Y(t) \sim p(X)p^{(s_t)}(Y(t) \mid X)}(Y(t) \in \widehat{C}_t(X)) \geq 1 - \alpha, \tag{12}$$

for any data generating distribution $\mathbb{P}_0$ with the corresponding selection score $s_t \in \mathcal{E}(\Gamma)$.

The goal is to construct a predictive interval that satisfies Definition 2 with $\mathcal{E}(\Gamma)$ defined by the MSM. Our first step is to create a valid predictive interval under a sensitivity model.

Coverage Guarantees for a Fixed Sensitivity Model. Given a fixed $s_t \in \mathcal{E}(\Gamma)$, plugging Equation (11) to Equation (3), the conformal weight $w_t^{(s_t)}(x, y)$ becomes a function of s_t. Let $w_i^{(s_t)} =$

$w_t^{(s_t)}(Z_i)$ be the conformal weight for Z_i, the predictive band in Equation (4) becomes

$$\widehat{C}^{(s_t)}(x) = \Big\{y \in \mathbb{R} : V(x,y) \le Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i^{(s_t)}\delta_{V_i} + p_{n+1}^{(s_t)}\delta_{\infty}\Big)\Big\}, \tag{13}$$

where $\{p_i^{(s_t)}\}_{i=1}^{n+1}$ normalizes $\{w_i^{(s_t)}\}_{i=1}^{n+1}$ and n is the number of data points.

In the following theorem, we show that $\widehat{C}^{(s_t)}(x)$ has a valid coverage given a specific sensitivity model s_t, when the propensity score is either known or estimable.

Lemma 3. Under SUTVA and strong overlap, for a selection score $s_t \in \mathcal{E}(\Gamma)$, we have

1. With a known propensity score $e(X)$, the predictive band in Equation (13) has coverage

$$1-\alpha \le \mathbb{P}_{X,Y(t)\sim p(X)p^{(s_t)}(Y(t)\,|\,X)}(Y(t) \in \widehat{C}_t^{(s_t)}(X)) \le 1-\alpha+\frac{\Gamma/\eta}{n+\Gamma/\eta}. \tag{14}$$

2. With an estimated propensity score $\widehat{e}(X)$, if $\eta < \widehat{e}(X_i) < 1-\eta$ almost surely for a constant $\eta \in (0, 0.5)$, the predictive band $\widehat{C}^{(s_t)}(x)$ in Equation (13) has a coverage probability

$$1-\alpha-\Delta \le \mathbb{P}(Y(t) \in \widehat{C}^{(s_t)}(X)) \le 1-\alpha+\frac{\Gamma/\eta}{n+\Gamma/\eta}+\Delta, \tag{15}$$

$$\Delta = \frac{\Gamma}{2}p(T=t)\mathbb{E}_{x\sim p(X\,|\,T=t)}\Big|\frac{1}{\widehat{e}(x)^t(1-\widehat{e}(x))^{1-t}} - \frac{1}{e(x)^t(1-e(x))^{1-t}}\Big|.$$

With a known propensity score, Equation (14) demonstrates that the coverage of the predictive band $\widehat{C}^{(s_t)}(x)$ is valid and is close to the nominal level. The closeness depends on the overlapping and confounding strength. When the estimated propensity score $\widehat{e}(X)$ differs from the true propensity score $e(X)$, as shown in Equation (15), the coverage probability might have an extra slack quantity Δ. The reason for $\widehat{e}(X) \neq e(X)$ could be the estimation error from the finite sample or the inference error from a misspecified treatment model.

Union Method. Based on the predictive band $\widehat{C}^{(s_t)}(x)$, we propose a union method that achieves the valid coverage under the MSM. That is, we now consider the worst-case coverage for all sensitivity models $s_t \in \mathcal{E}(\Gamma)$.

Proposition 1. Suppose the predictive interval $\widehat{C}^{(s_t)}(X) = [L^{(s_t)}(X), U^{(s_t)}(X)]$ satisfies

$$\mathbb{P}_{X,Y(t)\sim p(X)p^{(s_t)}(Y(t)\,|\,X)}(Y(t) \in \widehat{C}_t^{(s_t)}(X)) \ge 1-\alpha \tag{16}$$

for each $s_t \in \mathcal{E}(\Gamma)$. Then let $L = \inf_{s_t\in\mathcal{E}(\Gamma)} L^{(s_t)}$ and $U = \sup_{s_t\in\mathcal{E}(\Gamma)} U^{(s_t)}$, the interval $\widehat{C}_t^{\Gamma}(X) = [L,U] = \cup_{s_t\in\mathcal{E}(\Gamma)}[L^{(s_t)}, U^{(s_t)}]$ is a predictive interval for $Y(t)$ with at least $(1-\alpha)$ coverage under the sensitivity models $\mathcal{E}(\Gamma)$.

Proposition 1 states that to obtain a valid predictive interval for the MSM, we can first compute the predictive interval under a specific sensitivity model as in Equation (13), then take the union set by finding the extreme endpoints of such intervals over all the sensitivity models. By Equation (13), finding the extreme endpoints is equivalent to solving

$$\max_{s_t\in\mathcal{E}(\Gamma)} Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i^{(s_t)}(Z_{1:n},(X,y))\delta_{V_i} + p_{n+1}^{(s_t)}(Z_{1:n},(X,y))\delta_{\infty}\Big). \tag{17}$$

However, in practice, it is difficult to directly search over the sensitivity models in $\mathcal{E}(\Gamma)$, because the elements of $\mathcal{E}(\Gamma)$ are not defined parametrically.

Quantile Optimization with Linear Constraints. To operationalize Equation (17), we can search over the conformal weights instead of the sensitivity models. As Equation (17) shows, a sensitivity model influences the predictive band only through the conformal weights.

In Section 3.2, we find the range of conformal weights under an MSM. Denote the upper and lower bounds of the conformal weights in Lemma 2 as

$$w_{lo}^{\Gamma}(x) := \Big(1+\frac{1}{\Gamma}\Big(\frac{1-e(x)}{e(x)}\Big)^{2t-1}\Big)p(T=t),$$
$$w_{hi}^{\Gamma}(x) := \Big(1+\Gamma\Big(\frac{(1-e(x))}{e(x)}\Big)^{2t-1}\Big)p(T=t). \tag{18}$$

Then the optimization in Equation (17) simplifies to a constrained optimization problem,

$$\begin{aligned} \max_{w_{1:n+1}} \quad & Q_{1-\alpha}\Big(\textstyle\sum_{i=1}^{n} p_i\delta_{V_i} + p_{n+1}\delta_{\infty}\Big). \\ \text{subject to} \quad & p_i = \tfrac{w_i}{\sum_{i=1}^{n+1} w_i},\ 1 \le i \le n+1 \\ & w_{lo}^{\Gamma}(X_i) \le w_i \le w_{hi}^{\Gamma}(X_i),\ 1 \le i \le n, \\ & w_{lo}^{\Gamma}(X) \le w_{n+1} \le w_{hi}^{\Gamma}(X), \end{aligned} \tag{19}$$

where the conformal weights $\boldsymbol{w} = (w_1,\ldots,w_n,w_{n+1})$ are the optimizing variables. For notational convenience, we suppress the superscript t and denote $w_i = w(X_i,Y_i)$ for $1 \le i \le n$, and $w_{n+1} = w(X,y)$.

An efficient algorithm to solve Equation (19) can be designed by characterizing its optima.

Proposition 2. Without loss of generality, suppose $X_{1:n}$ are labeled such that the nonconformity scores are ordered, $V_1 \le V_2 \le \cdots \le V_n < V_{n+1} = \infty$, and let

$$\widehat{k} = \max\Big\{k \in [n+1] : \text{for } k \le j \le n,\ w_j = w_{hi}^{\Gamma}(X_j), \ w_{n+1} = w_{hi}(X),\ \sum_{j=k}^{n+1} p_j \ge \alpha; \ \text{for } j < k, w_j = w_{lo}^{\Gamma}(X_j)\Big\}.$$

Then the optima of Equation (19) is $\widehat{\boldsymbol{w}} = (w_{lo}^{\Gamma}(X_1),\ldots, w_{lo}^{\Gamma}(X_{\widehat{k}-1}), w_{hi}^{\Gamma}(X_{\widehat{k}}),\ldots,w_{hi}^{\Gamma}(X_n), w_{hi}^{\Gamma}(X))$. Furthermore, the optimal objective value $\widehat{Q}(Z_{1:n},X;\Gamma,\alpha,t) = V_{\widehat{k}}$.

According to Proposition 2, to solve Equation (19), we first sort $V_{1:n+1}$ in ascending order and initialize the conformal weights $w_i = w_{lo}^{\Gamma}(X_i)$ for $1 \le i \le n$, $w_{n+1} = w_{lo}^{\Gamma}(X)$. Then we iteratively flip w_k from $w_{lo}^{\Gamma}(X_k)$ to $w_{hi}^{\Gamma}(X_k)$ for $k = n+1, n, \ldots, 1$ until $\sum_{i=k}^{n+1} p_i \ge \alpha$. Supposing the iteration stops at $k = m$, the optimal objective value in Equation (19) is uniquely determined as $\widehat{Q}(Z_{1:n}, X; \Gamma, \alpha, t) = V_m$.

To sum up, by Lemma 3, the interval $\widehat{C}_t^{(s_t)}(x)$ in Equation (13) satisfies the coverage of Equations (14) and (15). With $\widehat{Q}$ given in Proposition 2,

$$\widehat{C}_t^{\Gamma}(x) = \left\{ y \in \mathbb{R} : V(x,y) \le \widehat{Q}(Z_{1:n}, x; \Gamma, \alpha, t) \right\} \tag{20}$$

is the union set $\cup_{s_t \in \mathcal{E}(\Gamma)} \widehat{C}^{(s_t)}(x)$. According to Proposition 1, $\widehat{C}_t^{\Gamma}(x)$ is a valid predictive interval of a missing counterfactual outcome under the MSM.

Theorem 1. Under the condition of Lemma 3, with known propensity score $e(X)$, the predictive band $\widehat{C}_t^{\Gamma}(x)$ in Equation (20) has nominal coverage $1-\alpha$ of $Y(t)$ under the collection of sensitivity models $\mathcal{E}(\Gamma)$; with an estimated $\widehat{e}(X)$, the coverage is at least to the lower bound in Equation (15).

Implementation and Computational Cost. We compute the predictive interval by adopting the framework of split conformal prediction (Papadopoulos et al. 2002; Lei and Wasserman 2014). Classic conformal inference fits the predictive function using the leave-one-out observed data to ensure the exchangeability and has to fit a predictor multiple times. Spilt conformal prediction reduces the computational cost by randomly splitting the observed data into a preliminary set and a calibration set. The prediction model is fitted on the preliminary set for one time, set as fixed, and used to compute the nonconformity scores on the calibration set and target set.

For CSA, the predictive interval in Equation (20) can be computed analytically on top of a specific conformal inference algorithm. As an example, for the split conformal inference with nonconformity score $V_i = |Y_i - \widehat{\mu}(X_i)|$ (Lei, Rinaldo, and Wasserman 2015), where $\widehat{\mu}(\cdot)$ is the mean response function, Equation (20) becomes $\widehat{C}_t^{\Gamma}(x) = [\widehat{\mu}(x) - \widehat{Q}(Z_{1:n}, x; \Gamma, \alpha),\ \widehat{\mu}(x) + \widehat{Q}(Z_{1:n}, x; \Gamma, \alpha)]$. For split conformal quantile regression with nonconformity score $V_i = \max\{\widehat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \widehat{q}_{1-\alpha/2}(X_i)\}$ (Romano, Patterson, and Candès 2019; Lei and Candès 2021), where $\widehat{q}(\cdot)$ is the conditional quantile function, Equation (20) becomes $\widehat{C}_t^{\Gamma}(x) = [\widehat{q}_{\alpha/2}(x) - \widehat{Q}(Z_{1:n}, x; \Gamma, \alpha),\ \widehat{q}_{1-\alpha/2}(x) + \widehat{Q}(Z_{1:n}, x; \Gamma, \alpha)]$. The full algorithm is summarized in Algorithm 1.

For each target unit, to solve the optimization in Equation (19), the computational complexity is $\mathcal{O}(mn)$ if the loop ends in m iterations and the worst-case complexity is $\mathcal{O}(n^2)$. When the target coverage $1-\alpha$ is high, m is close to 1 and the total computation time is close to the optimal rate that is needed to evaluate the objective function for one time. Other computations are sorting $V_{1:n+1}$ and fitting the treatment and outcome models, which can be shared by different target units. Therefore, CSA is highly efficient, inducing little extra computation comparing to the conformal prediction under unconfoundedness.

3.4. Predictive Band for the Individual Treatment Effect

We now develop a sensitivity analysis for the ITE of a target unit, for which both potential outcomes are unobserved. Let the covariates of a target unit be X. Using the data of the treatment group t, by Algorithm 1, we can construct an interval $\widehat{C}_t^{\Gamma}(X) = [L_t^{\Gamma}(X), U_t^{\Gamma}(X)]$ which has $1-\alpha_t$ coverage of $Y(t)$ under the sensitivity models $\mathcal{E}(\Gamma)$. Let $\widehat{C}^{\Gamma}(X) = [L_1^{\Gamma}(X) - U_0^{\Gamma}(X), U_1^{\Gamma}(X) - L_0^{\Gamma}(X)]$ and $\alpha_1 + \alpha_0 = \alpha$. By the Bonferroni correction,

$$\mathbb{P}(Y(1) - Y(0) \in \widehat{C}^{\Gamma}(X)) \ge 1 - \mathbb{P}(Y(1) \notin \widehat{C}_1^{\Gamma}(X) \text{ or } Y(0) \notin \widehat{C}_0^{\Gamma}(X)) \ge 1 - \alpha. \tag{21}$$

So the predictive interval $\widehat{C}^{\Gamma}(X)$ has the desired coverage. Though computationally simple, the Bonferroni method might produce overly conservative interval $\widehat{C}^{\Gamma}(X)$ for the ITE, because the coverage $1-\alpha_t$ for each potential outcome is higher than $1-\alpha$.

To mitigate this problem, we follow (Lei and Candès 2021) and develop a nested approach. The idea is to first randomly sample a subset of data as the validation set (indexed by $\mathcal{I}_{\text{val}}$) and set the rest of the observed data as the nonvalidation set. For each individual $i \in \mathcal{I}_{\text{val}}$, let $\widehat{C}^{\Gamma}(X_i) = \widehat{C}_1^{\Gamma}(X_i) - Y_i(0)$ if $T_i = 0$ and $\widehat{C}^{\Gamma}(X_i) = Y_i(1) - \widehat{C}_0^{\Gamma}(X_i)$ if $T_i = 1$. The coverage probability decomposes as

$$\begin{aligned}\mathbb{P}(Y_i(1) - Y_i(0) \in \widehat{C}^{\Gamma}(X_i)) &= \mathbb{P}(T_i = 1)\mathbb{P}(Y_i(0) \in \widehat{C}_0^{\Gamma}(X_i) | T_i = 1) \\ &\quad + \mathbb{P}(T_i = 0)\mathbb{P}(Y_i(1) \in \widehat{C}_1^{\Gamma}(X_i) | T_i = 0).\end{aligned}$$

If the interval $\widehat{C}_t^{\Gamma}(X_i)$ has a coverage probability of $Y_i(t)$ higher than $1-\alpha$, the coverage probability of $\widehat{C}^{\Gamma}(X)$ for the ITE is also higher than $1-\alpha$. The dataset $\widetilde{\mathcal{D}} = \{X_i, \widehat{C}^{\Gamma}(X_i)\}_{i \in \mathcal{I}_{\text{val}}}$ with $X_i \overset{\text{iid}}{\sim} p(X)$ can be used to fit a predictive function $X \mapsto \widehat{C}^{\Gamma}(X)$, which maps to a predictive interval $\widehat{C}^{\Gamma}(X)$ for a data point with covariates $X \sim p(X)$. The mapping can be two regressions with the input as X_i, $i \in \mathcal{I}_{\text{val}}$ and the output as the upper and lower endpoints of $\widehat{C}^{\Gamma}(X_i)$, respectively. This becomes a relatively easy in-sample prediction problem.

We use Algorithm 1 to obtain the predictive intervals $\widehat{C}_t^{\Gamma}(X_i)$ for the data points in the validation set. The training data are from the treatment group $1-t$ in the non-validation set and the target data are from the treatment group t in the validation set, $t = 0, 1$. Similar to Lemma 2, the bounds of the conformal weights $w_t(x, y)$ can be computed as

$$\frac{p(X \mid T = t)p(Y(t) \mid X, T = t)}{p(X \mid T = 1-t)p(Y(t) \mid X, T = 1-t)} \in \Big[\frac{1}{\Gamma}\Big(\frac{e(x)}{1-e(x)}\Big)^{2t-1}, \Gamma\Big(\frac{e(x)}{1-e(x)}\Big)^{2t-1}\Big]. \tag{22}$$

The algorithm is summarized in Algorithm 2.

3.5. Sharpness via Covariates Balancing

Notion of sharpness. Consider the sharpness on the sensitivity models. A sharp sensitivity model should be data compatible and not have observational implications (Franks, D'Amour, and Feller 2019; Dorn and Guo 2022). For the MSM, we define the sharp MSM as

Algorithm 2: CSA for the ITE estimation with Nested Method

Input: Data $(X_i, T_i, Y_i(T_i))_{i=1}^N$, level α, sensitivity parameter Γ, target covariates X

Step I: Preliminary processing

1: Split the data into 2-fold, indexed by $\mathcal{I}$ and $\mathcal{I}_{\text{val}}$

2: Denote the treated and control group data in $\mathcal{I}$ ($\mathcal{I}_{\text{val}}$) as $\mathcal{I}^t, \mathcal{I}^c$ ($\mathcal{I}^t_{\text{val}}, \mathcal{I}^c_{\text{val}}$), respectively

Step II: Predictive interval for the ITE τ_i at the target point

1: Run Algorithm 1 with data in $\mathcal{I}, \mathcal{I}_{\text{pre}} \cup \mathcal{I}_{\text{cal}} = \mathcal{I}^t$ and for each target point $i \in \mathcal{I}^c_{\text{val}}$; the bounds of weight $w^\Gamma_{lo}(x) = \widehat{e}(x)/(\Gamma(1-\widehat{e}(x)))$ and $w^\Gamma_{hi}(x) = \Gamma\widehat{e}(x)/(1-\widehat{e}(x))$; return $\widehat{C}^\Gamma_1(X_i)$

2: For $i \in \mathcal{I}^c_{\text{val}}$, compute $\widehat{C}^\Gamma(X_i) = \widehat{C}^\Gamma_1(X_i) - Y_i(0)$

3: Run Algorithm 1 with data in $\mathcal{I}, \mathcal{I}_{\text{pre}} \cup \mathcal{I}_{\text{cal}} = \mathcal{I}^c$ and for each target point $i \in \mathcal{I}^t_{\text{val}}$; the bounds of weight $w^\Gamma_{lo}(x) = (1-\widehat{e}(x))/(\Gamma\widehat{e}(x))$ and $w^\Gamma_{hi}(x) = \Gamma(1-\widehat{e}(x))/\widehat{e}(x)$; return $\widehat{C}^\Gamma_0(X_i)$

4: For $i \in \mathcal{I}^t_{\text{val}}$, compute $\widehat{C}^\Gamma(X_i) = Y_i(1) - \widehat{C}^\Gamma_0(X_i)$

5: Learn the predictive function $X \to \widehat{C}^\Gamma(X)$ with training data $\{X_i, \widehat{C}^\Gamma(X_i)\}_{i\in\mathcal{I}_{\text{val}}}$; predict $\widehat{C}^\Gamma(X)$ for the target data with the learned predictive function

Output: $\widehat{C}^\Gamma(X)$

$$\mathcal{E}^*(\Gamma) = \{s_t(x,y) \in \mathcal{E}(\Gamma) : \int p^{(s_t)} (Y(1) = y \,|\, X = x, T = 0)dy = 1\}, \quad (23)$$

where $\mathcal{E}(\Gamma)$ is defined in Equation (5) and $p^{(s_t)}(Y(1) = y \,|\, X, T = 0)$ is the induced counterfactual distribution in Equation (10). The sharp MSM is a subset of the selection scores in $\mathcal{E}(\Gamma)$ that induces proper counterfactual density. By Lemma 1, for example, $\mathcal{E}^*(\Gamma)$ excludes the selection scores with an odds ratio $\text{OR}(s_t(X, Y(t)), e(X))$ uniformly greater (or less than) one.

Recent work improves the sharpness of the MSM in estimating the ATE (Zhao, Small, and Bhattacharya 2019; Dorn and Guo 2022). Dorn and Guo (2022) shows that the selection score s_1 is data compatible if it satisfies the constraint $\mathbb{E}[\frac{T}{s_1(X,Y(1))} \,|\, X] = 1$ (the unobserved confounder in Dorn and Guo (2022) is replaced with $Y(1)$). This constraint is equivalent to the constraint in Equation (23) as shown in Proposition 3, Appendix B, supplementary materials. We consider estimating $Y(1)$ for simplicity and the discussion applies to $Y(0)$ similarly. The derivation and computation details of this section are presented in Appendix B, supplementary materials.

Sharpness by Covariates Balancing. The integral constraint in Equation (23) is easy to interpret but is infeasible to compute because we often only observe one outcome value for a given X. However, the constraint, equivalent to $\mathbb{E}[\frac{T}{s_1(X,Y(1))} \,|\, X] = 1$ by Proposition 3, indicates that for an arbitrary vector-valued function $g(X)$,

$$\mathbb{E}[\frac{g(X_i)T_i}{s_1(X_i, Y_i(1))}] = \mathbb{E}_{X_i}\big[g(X_i)\mathbb{E}[\frac{T_i}{s_1(X_i, Y_i(1))} \,|\, X_i]\big] = \mathbb{E}[g(X_i)]. \quad (24)$$

By enforcing the condition in Equation (24) with different covariates function $g(X)$, we can reduce $\mathcal{E}(\Gamma)$ close to $\mathcal{E}^*(\Gamma)$. Since Equation (24) holds similarly for the control group, it represents the covariate balancing between the treated and control group. This means encouraging the covariate balancing improves the sharpness of the MSM.

We incorporate the balancing condition Equation (24) to CSA. By Lemma 1, we transform the condition in Equation (24) to the constraints in the quantile optimization in Equation (19). Specifically, we optimize Equation (19) with additional constraints

$$\frac{1}{N_t}\sum_{i:T_i=t} g_k(X_i)w_i^\Gamma = \frac{1}{N}\sum_{i=1}^N \frac{T_i^t(1-T_i)^{1-t}}{\widehat{e}(X_i)^t(1-\widehat{e}(X_i))^{1-t}} g_k(X_i), \quad 1 \le k \le K, \quad (25)$$

where $g_k(X)$, $k \in \{1, 2, \ldots, K\}$ are the balancing functions specified by the researcher. We call this algorithm conformalized sharp sensitivity analysis (CSSA) and summarize it in Algorithm 3, Appendix B, supplementary materials. The optima of Equation (19) with additional constraints Equation (25) is smaller than that of Equation (19), thereby reducing the size of predictive band in estimating the ITE. For example, we find choosing $g(X) = \widehat{e}(X)$ effectively improves the sharpness in simulations. Including additional balancing functions such as the quantile function of the outcome distribution (Dorn and Guo 2022), the identity function, and the derivative of the estimated propensity score (Imai and Ratkovic 2014) may further reduce $\mathcal{E}(\Gamma)$ to $\mathcal{E}^*(\Gamma)$.

Though the sharpness is necessary for claiming a causal estimate to be sensitive to unmeasured confounding, sensitivity analysis is often applied to corroborate a nonzero causal effect identified in the primary analysis. From this perspective, the deviation from the sharpness might be interpreted as conservativeness (Ding and VanderWeele 2016; Cinelli and Hazlett 2020; Veitch and Zaveri 2020). For the ITE estimation, such conservativeness increases our confidence that a positive (or negative) ITE is indeed robust to unmeasured confounding when the sensitivity analysis suggests so.

4. Practical Considerations of the Algorithms

We now discuss how to interpret the coverage probability of CSA, choose the conformal inference algorithms, calibrate the sensitivity parameter, and evaluate the ITE estimation.

Marginal and Conditional Coverage. The probability in the coverage statement of CSA is over both the covariates and the outcomes. Hence, the coverage guarantee should be interpreted in a marginal way instead of a conditional way. In other words, suppose the estimand τ is either the missing potential outcome $Y(t)$ or the ITE $Y(1) - Y(0)$. $\mathbb{P}(\tau \in \widehat{C}(X))$ means that if we construct a predictive band $\widehat{C}(X)$ for a unit randomly sampled as the target, the probability that $\widehat{C}(X)$ captures τ is at least $1-\alpha$. The randomness is over both the covariates and the potential outcomes.

The marginal coverage measures the quality of prediction *averaged* over the target units. It does not guarantee the coverage of τ for a given *fixed* target unit. The loss of conditional coverage

is unavoidable if no distributional assumption is imposed on the observed data (Barber et al. 2021a). However, it is possible for an algorithm with marginal coverage to achieve conditional coverage asymptotically, under additional regularization conditions, or to satisfy a relaxed conditional coverage definition (Barber et al. 2021a; Lei and Candès 2021). We refer to Barber et al. (2021a) for a detailed discussion on the definitions, limitations, and connections of different types of coverages.

Choice of Conformal Inference Methods. CSA is compatible with a variety of conformal inference algorithms. The main components of a conformal inference are the prediction model and the corresponding nonconformity score (Angelopoulos and Bates 2021). The choice of conformal inference algorithm hinges on the properties of the underlying outcome distribution, such as the homogeneity, skewness and multimodality. A good choice of inference method leads to high interpretability and small predictive interval. For example, to estimate the ITE, the predictive band as a single interval may be more interpretable than a nonconvex set (Sesia and Romano 2021); an algorithm capturing individual heterogeneity might produce a shorter interval and more informative estimate (Romano, Patterson, and Candès 2019). Nevertheless, the coverage validity of the predictive band does not depend on the choice of conformal inference methods.

Calibration of the Sensitivity Parameter. In the MSM, the sensitivity parameter Γ quantifies the confounding strength. While setting Γ to a proper value requires domain knowledge, the observed data can provide useful reference (Imbens 2003; Hsu and Small 2013; Kallus, Mao, and Zhou 2019). In the definition of MSM, Γ measures the effect of knowing a potential outcome on the treatment assignment. We can view the potential outcome as a type of covariate (Robins, Rotnitzky, and Scharfstein 2000) and compute the effect of an observed covariate on the treatment assignment. Specifically, we compute $\Gamma_{ij} = \mathrm{OR}(e(X_i), e((X_{\setminus j})_i))$ as the effect of the jth covariates on the treatment assignment of the ith unit, where $e(X_{\setminus j})$ is the propensity score estimated without the jth covariates. The domain experts can assess a plausible magnitude of Γ by referring to the magnitude of $\{\Gamma_{ij}\}_{i,j}$. Here, approximating $e(X_i)$ by $e((X_{\setminus j})_i))$ may introduce conservativeness to the estimated confounding strength. We refer to (Cinelli and Hazlett 2020; Veitch and Zaveri 2020) for a discussion on this issue.

Evaluating the Predictive Band of an ITE. When evaluating the ITE estimation by simulations, we need to sample the true ITEs, which requires generating random samples of all the potential outcomes. To generate $Y_i(t) \sim p(Y(t) \mid X_i)$, we can sample $T_i \sim \mathrm{Bern}(e(X_i))$ and $Y_i(t) \sim p(Y(t) \mid X_i, T_i)$. However, when $T_i = 1 - t$, generating $Y_i(t)$ depends on a sensitivity model, which is not defined parametrically in the MSM. To solve this problem, we propose a rejection sampling method to generate counterfactual samples. The details of this sampling method is presented in Appendix C, supplementary materials.

5. Empirical Studies

In this section, we answer the following questions using synthetic data: can CSA provide a desired coverage? Are the predictive intervals of the ITE overly conservative? How do the predictive intervals of ITE compare to the interval estimates of population-level causal estimands? Finally, we illustrate how to apply CSA in an observational study.

5.1. CSA for Estimating Counterfactual Outcome

Following the synthetic data generation in Lei and Candès (2021) and Wager and Athey (2018), the potential outcome $Y_i(1)$ is from

$$Y_i(1) = \mathbb{E}[Y_i(1) \mid T_i = 1, X_i] + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$
$$\mathbb{E}[Y_i(1) \mid T_i = 1, X_i] = f(X_{i1}) f(X_{i2}),\ f(x) = \frac{2}{1+\exp(-5(x-0.5))}, \tag{26}$$

where the covariates $X_i = (X_{i1}, \ldots, X_{id})^\top$, $X_{ij} \sim \mathrm{Unif}(0,1)$. The propensity score is $e(X_i) = \frac{1}{4}(1 + \beta_{2,4}(1 - X_{i1}))$, where $\beta_{2,4}$ is the CDF of beta distribution with parameters $(2,4)$. We generate $n = 3000$ training data points $(X_i, Y_i(1))_{i:T_i=1}$ with dimension of covariates $d = 20$. We take 75% of the training data as the preliminary set and the rest as the calibration set. For a calibration set with n_{cal} data points, the coverage probability of a new target data follows distribution $\mathrm{Beta}(n_{\mathrm{cal}} + 1 - \lfloor (n_{\mathrm{cal}} + 1)\alpha \rfloor, \lfloor (n_{\mathrm{cal}} + 1)\alpha \rfloor)$ (Angelopoulos and Bates 2021). We set the nominal level $1 - \alpha = 0.8$ in this simulation. We consider two settings: in the homoscedastic case, $\sigma \equiv 1$, and in the heteroscedastic case, $\sigma \sim \mathrm{Unif}(0.5, 1.5)$.

For CSA, we use the split conformal prediction with mean prediction (Papadopoulos et al. 2002; Lei, Rinaldo, and Wasserman 2015) and conformal quantile regression Romano, Patterson, and Candès (2019), denoted as CSA-M and CSA-Q, respectively. We implement CSSA in Section 3.5 with mean prediction, denoted as CSSA-M and set the balancing constraints $g(X_i)$ in Equation (25) as the estimated propensity score $\widehat{e}(X_i)$. We report the ITE estimated under no unobserved confounding (NUC) as a benchmark, denoted as ITE-NUC (Lei and Candès 2021). For all conformal inference methods, we use the random forest (Breiman 2001) as the regression function.

We first assume the baseline outcome $Y(0) \equiv 0$. The estimation of ITE then reduces to estimating a single potential outcome $Y(1)$. Figure 1 demonstrates several counterfactual distributions $p(Y(1)|X, T = 0)$ that are generated by the rejection sampling method described in Section 4 with sensitivity models in $\mathcal{E}(\Gamma)$, $\Gamma = 4$. Unmeasured confounding is reflected as the difference between $p(Y(1)|X, T = 0)$ and $p(Y(1)|X, T = 1)$. Figure 1 show that the nonparametric MSM probes a variety of potential violations to unconfoundedness. We find the counterfactual distribution in the middle of Figure 1 results in the lowest coverage among the counterfactual cases in Figure 1 due to the mismatch of the high density regions between the observed and counterfactual distributions. Since the interval estimate by CSA has coverage guarantee for any sensitivity model in $\mathcal{E}(\Gamma)$, we report the results with counterfactual in the middle of Figure 1 as an adversarial case to test the validity of CSA.

In Appendix E Table 2, supplementary materials, we compare the interval estimates of CSA with those by a sensitivity analysis of the ATE (Zhao, Small, and Bhattacharya 2019) and an estimation of the CATE (Chipman, George, and McCulloch 2010). The (sub)population-based interval estimates underesti-

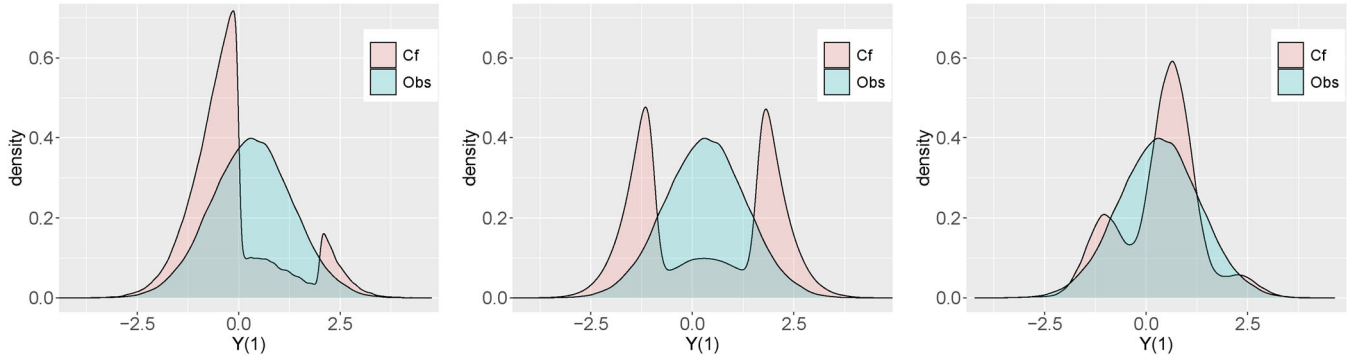


Figure 1. Distribution of $Y(1)$ for the synthetic data. Given the covariates X, Obs denotes the distribution $p(Y(1) \mid T = 1, X)$ in the observed group. Cf denotes the distribution of counterfactual outcome $p(Y(1) \mid T = 0, X)$. The three plots correspond to different sensitivity models in $\mathcal{E}(\Gamma)$.

mate the individual-level uncertainty and undercover the true ITE. This validates the necessity of individual-level sensitivity analysis.

Figure 2 illustrates the properties of the estimates by CSA and CSSA. The top panels show the empirical coverage under different confounding strengths. The empirical coverage is computed as $(\sum_{i=1}^{m} \mathbf{1}[Y_i(1) \in \widehat{C}_1^{\Gamma}(X_i)])/m$ for $m = 10{,}000$ target points. We observe that ITE-NUC achieves the target coverage under unconfoundedness ($\Gamma = 1$) but its coverage decreases as the confounding strength increases. In contrast, CSA-M and CSA-Q have valid coverage across all levels of Γ. The coverage of CSSA-M is above the nominal level and is lower than that of CSA-M, which demonstrate its validity and sharpness. For $\Gamma \leq 2.5$, CSSA-M has coverage centered at the nominal level which suggests its sharpness.

The middle panels in Figure 2 show the average interval length on the target units. We observe that the length of ITE-NUC remains the same as Γ changes. In comparison, the length of CSA and CSSA methods scales up with Γ, reflecting an increased uncertainty with stronger unmeasured confounding. On average, CSA-M produces shorter intervals than CSA-Q when the data is homoscedastic, and they have similar interval lengths when the data are heteroscedastic. CSSA-M creates shorted interval than CSA-M for all $\Gamma > 1$.

To further analyze the sharpness of CSA prediction, we manually shrink the length of the predictive intervals by a constant factor for all the units and keep the interval centers unchanged. The maximum shrinkage factor without losing the target coverage reflects the sharpness. From the bottom panels in Figure 2, we observe that the empirical coverage drops below the $1 - \alpha$ level if the shrink factor is above 10% and 15% for homoscedastic and heteroscedastic data, respectively. The maximal shrinkage factor being low means CSA methods produce relatively sharp intervals.

In Figure 3, we visualize the ITE estimates for multiple individuals. For each unit i, we compute the difference between the predictive interval and the true ITE as $\widehat{C}^{\Gamma}(X_i) - \tau_i$ which contains 0 if and only if $\widehat{C}^{\Gamma}(X_i)$ contains τ_i. For each method, we consider two confounding strengths $\Gamma \in \{1, 3\}$, set the coverage $1 - \alpha = 0.8$, and randomly sample 70 units. When there is unmeasured confounding, ITE-NUC produces a large fraction of intervals that do not contain the ITE, but CSA methods have a small fraction of undercovered intervals on average (less than $\alpha = 0.2$) for both confounding strengths.

5.2. CSA for the ITE Estimation

We further study when both $Y_i(1)$ and $Y_i(0)$ of a unit are unobserved. The outcome $Y(1)$ is generated according to Equation (26) and the observed outcome $Y(0)$ is generated by

$$\begin{aligned} Y_i(0) &= \mathbb{E}[Y_i(0) \mid T_i = 0, X_i] + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2); \\ \mathbb{E}[Y_i(0) \mid T_i = 0, X_i] &= f(X_{i1}) f(X_{i2}) + \tfrac{10 \sin(X_{i3})}{1 + \exp(-5X_{i3})}, \end{aligned} \tag{27}$$

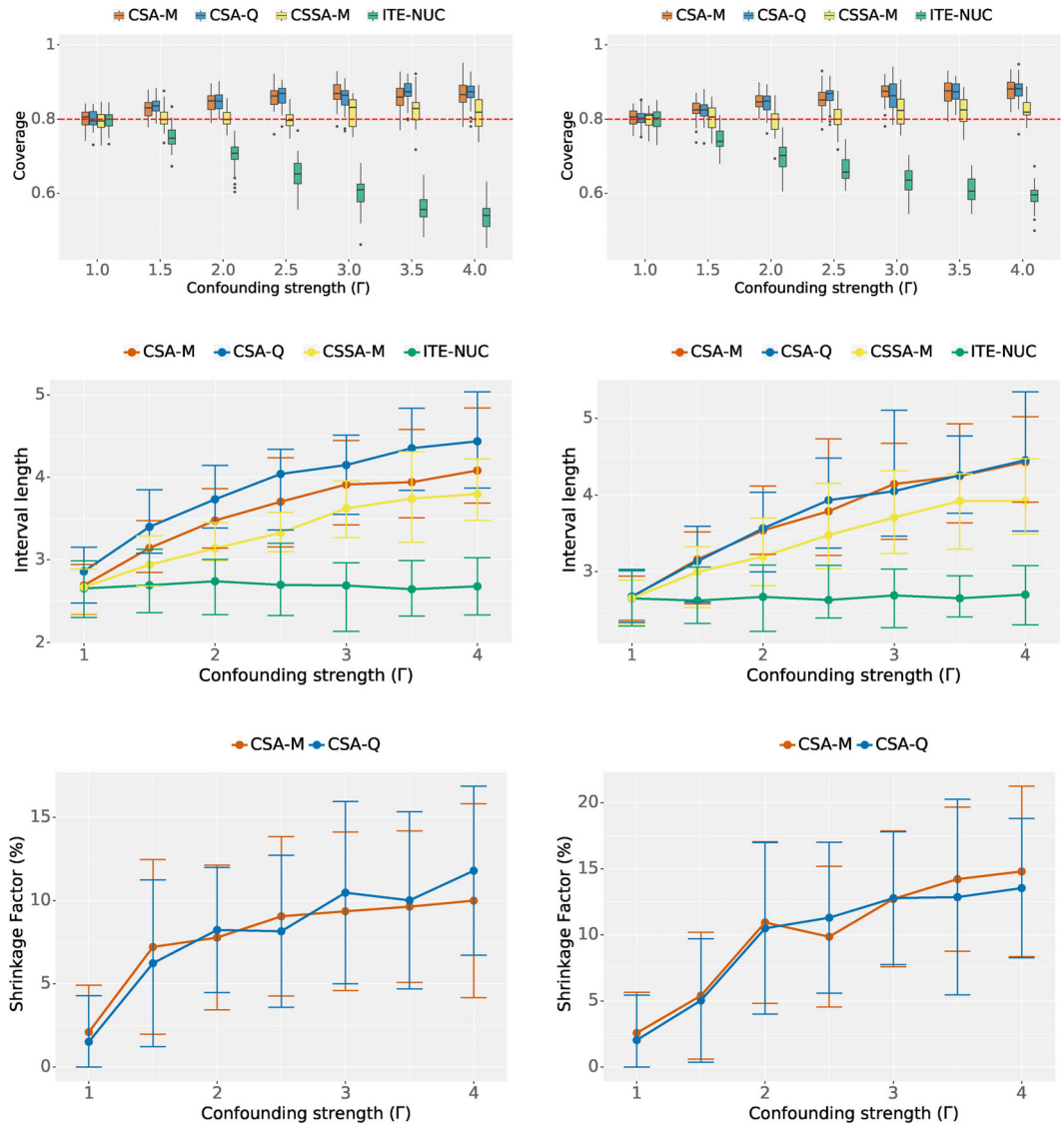
where $f(x)$ follows the definition in Equation (26). The construction of the counterfactual distribution $p(Y(0) | X = x, T = 0)$ is similar to the single missing outcome case, the details of which are in Appendix E, supplementary materials. We analyze the Bonferroni correction and the nested approach wrapped around CSA-M, CSA-Q and ITE-NUC. For the nested approach, following Lei and Candès (2021), we learn the mapping $X \mapsto \widehat{C}^{\Gamma}(X)$ in Section 3.4 by fitting 40% quantile of the lower endpoint and 60% quantile of the upper endpoint with quantile forest function in R package `grf`.

The results of coverage and interval length are shown in Figure 7, Appendix E, supplementary materials. The Bonferroni correction provides conservative interval estimates. In comparison, the interval estimation of CSA methods with the nested method are less conservative. Similar to Section 5.1, ITE-NUC has poor coverage when the unconfoundedness is violated, but CSA methods have valid coverages across different levels of confounding strength.

5.3. Application: ITEs of Fish Consumption on Blood Mercury

Finally, we illustrate the application of CSA using survey responses from the National Health and Nutrition Examination Survey (NHANES) 2013–2014. The causal question we study is the effect of high fish consumption on individuals' blood mercury levels when there is potentially unmeasured confounding.

Following Zhao, Small, and Bhattacharya (2019), we define the high fish consumption as more than 12 servings of fish a person consumes in the previous month and low fish consumption as 0 or 1 serving of fish. The outcome of interest is the blood mercury level, which is measured in ug/L and transformed to the logarithmic scale. The dataset contains $n = 1107$ units, where 80% are randomly sampled as training data and the rest 20% are the target units. There are $p = 8$ covariates about the demographics and health conditions (Zhao, Small,



(a) Homosc. (b) Heterosc.

Figure 2. *Top*: Empirical coverage of the ITE. Dashed line denotes the target coverage level. Across Γ, the predictive intervals produced by CSA and CSSA methods reach the valid coverage. CSSA-M improves sharpness over CSA-M. *Middle*: The average length of the predictive intervals. The interval lengths by CSA and CSSA methods increase with Γ, reflecting increased uncertainty under unmeasured confounding. *Bottom*: The sharpness of CSA. The maximal shrinkage factor that preserves the nominal coverage is low, which suggests CSA methods are relatively sharp. The error bar is by 100 independent trials.

and Rosenbaum 2018). We use random forest and quantile forest to fit the observed outcome, the gradient boosting to estimate the propensity score and the nested method with quantile forest as the interval prediction function.

We calibrate the sensitivity parameter Γ with the observed data. As discussed in Section 4, we compute Γ_{ij} as the effect of jth covariate on the treatment assignment of the ith unit in terms of odds ratio. Figure 4(a) shows the distribution of $\{\widetilde{\Gamma}_{ij}\}_{i=1:n}^{j=1:p}$ where $\widetilde{\Gamma}_{ij}$ equals to Γ_{ij} if $\Gamma_{ij} \geq 1$ and $1/\Gamma_{ij}$ otherwise. By Figure 4(a), we may consider $\Gamma \in [1, 3]$ as a plausible range of confounding strength. The choice of a proper sensitivity parameter often needs further domain knowledge in addition to the reference information from data.

For each target unit k, CSA produces an interval estimation $\widehat{C}^{\Gamma}(X_k) = [l_k, u_k]$. We call $\widehat{C}^{\Gamma}(X_k)$ a positive interval if $l_k > 0$, which represents a positive individual effect, and call $\widehat{C}^{\Gamma}(X_k)$ a negative interval if $u_k < 0$. Figure 4 reports the fraction of positive and negative intervals in the target units against the target coverage $1 - \alpha$ and the sensitivity parameter Γ. Overall, the fraction of positive intervals increases when the confounding strength and the target coverage decrease. There is a relatively strong evidence of positive effects when $\alpha \leq 0.2$ and $\Gamma \leq 2$, and there is no evidence of negative effects for $\alpha \leq 0.5$ and $\Gamma \leq 3$.

The results of individual-level estimates are reported in Figure 5. We randomly sample 70 individuals in the target set and show the predictive intervals of their ITEs with target coverage $1-\alpha = 0.8$. In Figure 5, the interval prediction of the treatment effect is heterogeneous across individuals. Under unconfoundedness, the ITEs for most individuals are likely to be positive. When $\Gamma = 2$, the effects of fish consumption for some individuals are explained away by the unmeasured confounding. From

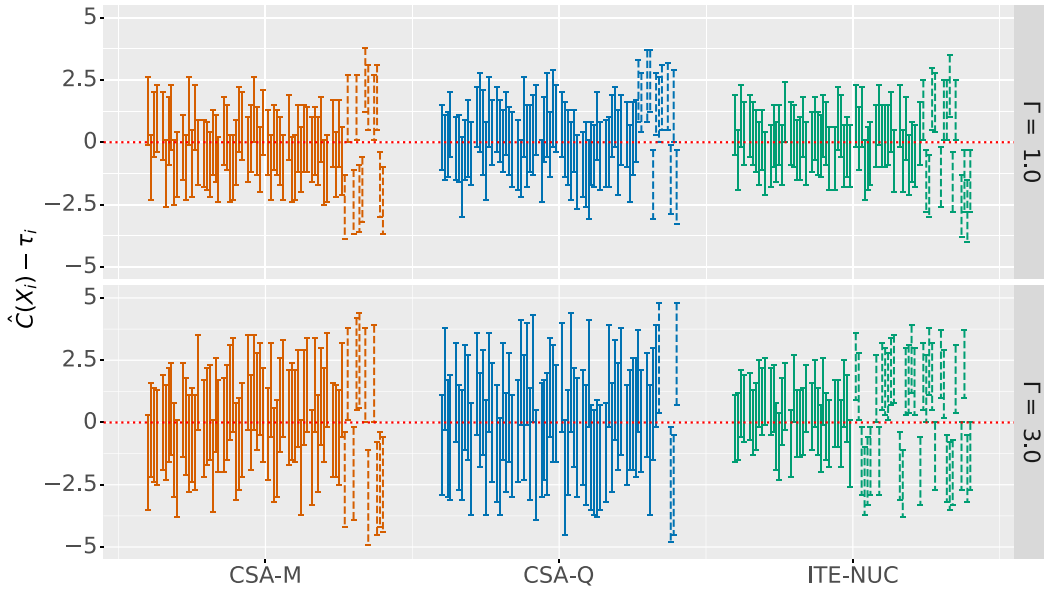


Figure 3. The figure reports predictive intervals for random individuals with confounding strengths. Each interval is the predictive interval minus the true ITE for one individual. Solid intervals contain 0 and dashed intervals do not contain 0. When $\Gamma = 1$, all methods have similar coverage at $1 - \alpha = 0.8$; when $\Gamma = 3$, ITE-NUC has high miscoverage while CSA maintains a valid coverage.

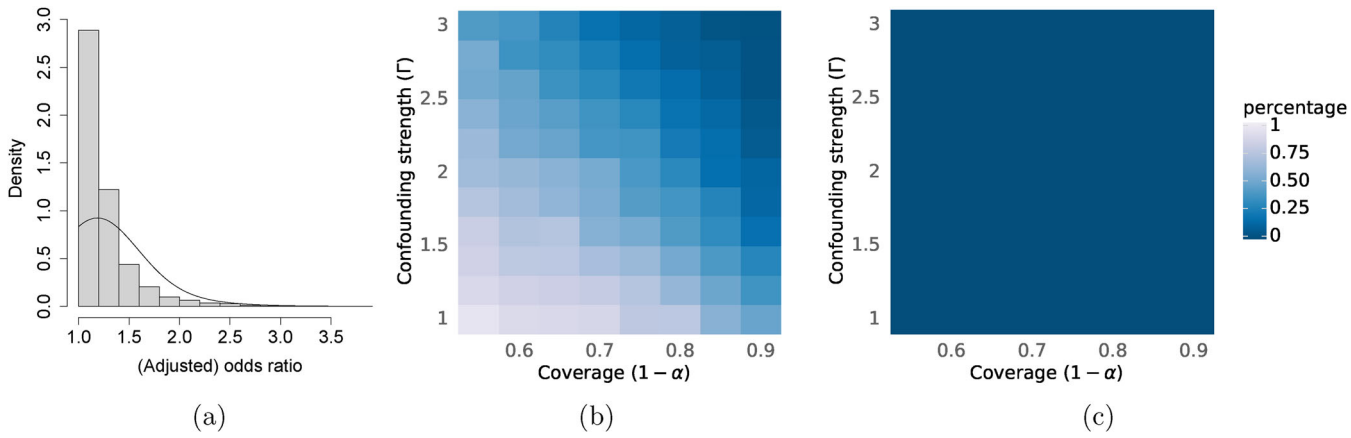


Figure 4. When the confounding strength is within the range of the study, we can say with high confidence that for a group of individuals, high fish consumption increases their blood mercury levels. The figures are produced using the NHANES fish consumption data. The predictive intervals are estimated by CSA-M. (a) provides reference information for the magnitude of the sensitivity parameter Γ from the observed covariates. (b) shows the fraction of intervals with positive lower bounds; (c) shows the fraction of intervals with negative upper bounds.

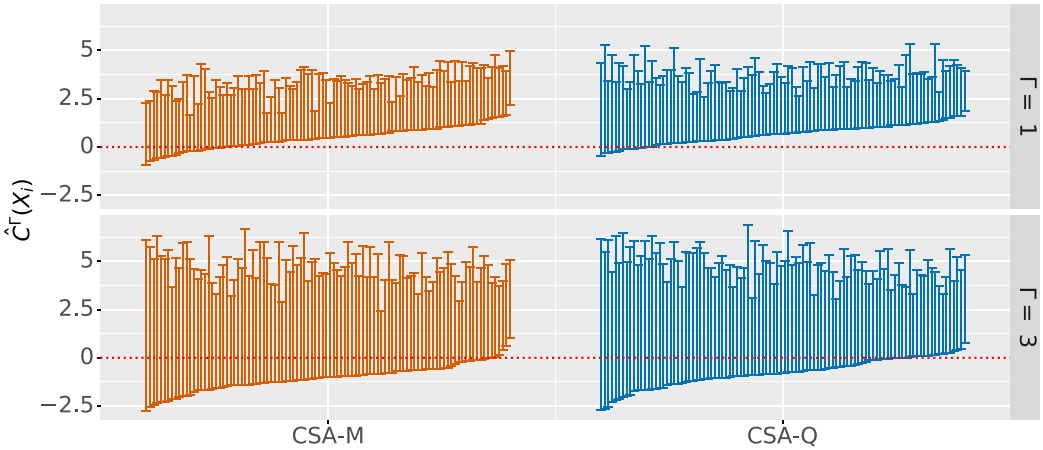


Figure 5. Predictive intervals for the target individuals with different unmeasured confounding strength. In top panels, when $\Gamma = 1$, a large fraction of individuals have positive effects. In the bottom panels, when $\Gamma = 2$, we can still identify individuals whose effects remain positive.

Figure 5, we can tell the subgroup for whom the effect of fish consumption on the blood mercury level is relatively insensitive to the unmeasured confounding. The predictive intervals given by the sensitivity analysis can thus provide useful information to guide personal decisions on fish consumption.

6. Discussion

In this article we developed a sensitivity analysis method for the ITE called CSA. We developed CSA by extending conformal inference to distribution shift. We adopted a two-stage design to propagate the uncertainty of an unmeasured confounding to the predictive interval of the ITE. We provided theoretical guarantees on the coverage property of the predictive interval, designed CSSA to improve the sharpness of CSA, and developed a rejection sampling method to evaluate the performance in simulation. Finally, we analyze CSA and CSSA using synthetic data and demonstrate the application in an observational study.

There are many directions for future research. We quantified the confounding strength by the MSM. Further research could explore alternative types of sensitivity models. If the nature of confounding is known, it might be preferable to model the effect of a confounder parametrically. We can also make a sensitivity assumption on the dependency structure between potential outcomes, which may improve the sharpness. Such dependencies can, for example, be modeled by a copula (Franks, D'Amour, and Feller 2019; Zheng, D'Amour, and Franks 2021). Finally, the extended conformal prediction might be used to test other untestable assumptions, such as the invariant causal mechanism (Peters, Bühlmann, and Meinshausen 2016).

Supplementary Materials

The Supplementary Material contains the technical proofs of the theoretical results, further discussions on the algorithms and practical considerations, and detailed results of the empirical studies.

Acknowledgments

The authors thank the Editor, Associate Editor, and three anonymous reviewers for helpful and constructive comments. The authors thank Gemma Moran and Simon Tavaré for their valuable feedback. Mingzhang Yin thank the support from Data Science Institute and Irving Institute for Cancer Dynamics, Columbia University.

Funding

This work is supported by NSF IIS 2127869, ONR N00014-17-1-2131, ONR N00014-15-1-2209, Simons Foundation, Sloan Foundation, Open Philanthropy.

ORCID

Mingzhang Yin http://orcid.org/0000-0002-5216-2437

References

Amin-Chowdhury, Z., and Ladhani, S. N. (2021), "Causation or Confounding: Why Controls are Critical for Characterizing Long Covid," *Nature Medicine*, 27, 1129–1130. [1]

Angelopoulos, A. N., and Bates, S. (2021), "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," arXiv no. 2107.07511. [9]

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021a), "The Limits of Distribution-Free Conditional Predictive Inference," *Inf inference*, 10, 455–482. [9]

——— (2021b), "Predictive Inference with the Jackknife+," *Annals of Statistics*, 49, 486–507. [3]

Blackwell, M. (2014), "A Selection Bias Approach to Sensitivity Analysis for Causal Effects," *Political Analysis*, 22, 169–182. [2]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [9]

Brook, D. (1964), "On the Distinction between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems," *Biometrika*, 51, 481–483. [5]

Brumback, B. A., Hernán, M. A., Haneuse, S. J., and Robins, J. M. (2004), "Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures," *Statistics in Medicine*, 23, 749–767. [2]

Candès, E. J., Lei, L., and Ren, Z. (2021), "Conformalized Survival Analysis," arXiv:2103.09763. [2]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *Annals of Applied Statistics*, 4, 266–298. [9]

Cinelli, C., and Hazlett, C. (2020), "Making Sense of Sensitivity: Extending Omitted Variable Bias," *Journal of the Royal Statistical Society*, Series B, 82, 39–67. [2,8,9]

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203. [2]

Ding, P., and VanderWeele, T. J. (2016), "Sensitivity Analysis Without Assumptions," *Epidemiology (Cambridge, Mass.)*, 27, 368–377. [2,8]

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016), "A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding," *Statistics in Medicine*, 35, 3453–3470. [2]

Dorn, J., and Guo, K. (2022), "Sharp Sensitivity Analysis for Inverse Propensity Weighting via Quantile Balancing," *Journal of the American Statistical Association*, 1–8. [7,8]

Franks, A., D'Amour, A., and Feller, A. (2019), "Flexible Sensitivity Analysis for Observational Studies Without Observable Implications," *Journal of the American Statistical Association*, 115, 1730–1746. [2,5,7,13]

Franks, A. M., Airoldi, E. M., and Rubin, D. B. (2016), "Non-standard Conditionally Specified Models for Non-Ignorable Missing Data," arXiv (1603.06045). [5]

Greenland, S., Pearl, J., and Robins, J. M. (1999), "Confounding and Collapsibility in Causal Inference," *Statistical Science*, 14, 29–46. [1]

Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., Brooks, N., Smaja, M., Mircus, G., Pan, K., Southern, J., Swerdlow, D. L., Jodar, L., Levy, Y., and Alroy-Preis, S. (2021), "Impact and Effectiveness of mrna bnt162b2 Vaccine Against sars-cov-2 Infections and covid-19 Cases, Hospitalisations, and Deaths Following a Nationwide Vaccination Campaign in Israel: An Observational Study Using National Surveillance Data," *The Lancet*, 397, 1819–1829. [1]

Hernan, M. A., and Robins, J. M. (2010), *Causal Inference*, Boca Raton, FL: CRC. arXiv no. 2105.14045. [3]

Hoff, P. (2021), "Bayes-Optimal Prediction with Frequentist Coverage Control," arXiv no. 2105.14045. [2]

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960. [3]

Hong, G., Yang, F., and Qin, X. (2021), "Did you Conduct a Sensitivity Analysis? A New Weighting-based Approach for Evaluations of the Average Treatment Effect for the Treated," *Journal of the Royal Statistical Society*, Series A, 184, 227–254. [2]

Hsu, J. Y., and Small, D. S. (2013), "Calibrating Sensitivity Analyses to Observed Covariates in Observational Studies," *Biometrics*, 69, 803–811. [9]

Imai, K., and Ratkovic, M. (2014), "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society*, Series B, 76, 243–263. [8]

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132. [2,9]

Izbicki, R., Shimizu, G., and Stern, R. (2020), "Flexible Distribution-Free Conditional Predictive Bands Using Density Estimators," in *AISTATS*. [2]

Jesson, A., Mindermann, S., Gal, Y., and Shalit, U. (2021), "Quantifying Ignorance in Individual-Level Causal-Effect Estimates Under Hidden Confounding," arXiv:2103.04850. [2]

Jin, Y., Ren, Z., and Candès, E. J. (2021), "Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach." arXiv:2111.12161. [2]

Kallus, N., Mao, X., and Zhou, A. (2019), "Interval Estimation of Individual-Level Causal Effects Under Unobserved Confounding," in *AISTATS*. [2,9]

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020a), "Adaptive, Distribution-Free Prediction Intervals for Deep Networks," in *AISTATS*. [2]

Kivaranovic, D., Ristl, R., Posch, M., and Leeb, H. (2020b), "Conformal Prediction Intervals for the Individual Treatment Effect," arXiv:2006.01474. [1,3]

Lei, J., and Wasserman, L. (2014), "Distribution-Free Prediction Bands for Non-parametric Regression," *Journal of the Royal Statistical Society*, Series B, 76, 71–96. [2,7]

Lei, J., Rinaldo, A., and Wasserman, L. (2015), "A Conformal Prediction Approach to Explore Functional Data," *Annals of Mathematics and Artificial Intelligence*, 74, 29–43. [2,7,9]

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111. [2]

Lei, L., and Candès, E. (2021), "Conformal Inference of Counterfactuals and Individual Treatment Effects," *Journal of the Royal Statistical Society*, Series B, 83, 911–938. [1,3,4,7,9,10]

Neyman, J. (1923), "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51. [3]

Papadopoulos, H. (2008), *Inductive Conformal Prediction: Theory and Application to Neural Networks*, Rijeka: INTECH Open Access Publisher. [2]

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002), "Inductive Confidence Machines for Regression," in *ECML*. [2,7,9]

Peters, J., Bühlmann, P., and Meinshausen, N. (2016), "Causal Inference by using Invariant Prediction: Identification and Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 78, 947–1012. [13]

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, eds. M. E. Halloran and D. Berry, pp. 1–94, New York: Springer. [2,4,9]

Romano, Y., Patterson, E., and Candès, E. (2019), "Conformalized Quantile Regression," *Advances in neural information processing systems*, 32. [2,7,9]

Rosenbaum, P. R., and Rubin, D. B. (1983a), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Series B, 45, 212–218. [2]

——— (1983b), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [3,4]

Rosenbaum, P. R. (2002), *Observational Studies*, (2nd ed.), New York: Springer. [1]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [3]

——— (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," *Journal of the American Statistical Association*, 75, 591–593. [3]

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120. [4]

Sesia, M., and Candès, E. (2020), "A Comparison of Some Conformal Quantile Regression Methods," *Stat*, 9, e261. [2]

Sesia, M., and Romano, Y. (2021), "Conformal Prediction Using Conditional Histograms," *Advances in Neural Information Processing Systems*, 34, 6304–6315. [2,9]

Shafer, G., and Vovk, V. (2008), "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, 9, 371–421. [2]

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637. [1,2,4]

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019), "Conformal Prediction under Covariate Shift," in *NeurIPS*. [1,2,3,4]

Veitch, V., and Zaveri, A. (2020), "Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias due to Unobserved Confounding," in *NeurIPS*. [2,8,9]

Vovk, V. (2012), "Conditional Validity of Inductive Conformal Predictors," in *Asian Conference on Machine Learning*. [2]

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic Learning in a Random World*, New York: Springer. [1,2,3,4]

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009), "On-Line Predictive Linear Regression," *Annals of Statistics*, 37, 1566–1590. [2]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [9]

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018), "Bounds on the Conditional and Average Treatment Effect with Unobserved Confounding Factors," arXiv:1808.09521. [2]

Zhao, Q., Small, D., and Rosenbaum, P. (2018), "Cross-Screening in Observational Studies that Test Many Hypotheses," *Journal of the American Statistical Association*, 113, 1070–1084. [11]

Zhao, Q., Small, D., and Bhattacharya, B. (2019), "Sensitivity Analysis for Inverse Probability Weighting Estimators via the Percentile Bootstrap," *Journal of the Royal Statistical Society*, Series B, 81, 735–761. [2,4,8,9,10]

Zheng, J., D'Amour, A., and Franks, A. (2021), "Copula-based Sensitivity Analysis for Multi-Treatment Causal Inference with Unobserved Confounding," arXiv:2102.09412. [13]