Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: amstat.tandfonline.com/journals/uasa20

Distribution-Free Prediction Intervals Under Covariate Shift, With an Application to Causal Inference

Jing Qin, Yukun Liu, Moming Li & Chiung-Yu Huang

To cite this article: Jing Qin, Yukun Liu, Moming Li & Chiung-Yu Huang (12 Jun 2024): Distribution-Free Prediction Intervals Under Covariate Shift, With an Application to Causal Inference, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2356886

To link to this article: https://doi.org/10.1080/01621459.2024.2356886

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2024.2356886

Taylor & Francis
Taylor & Francis Group

Distribution-Free Prediction Intervals Under Covariate Shift, With an Application to Causal Inference

Jing Qin[a], Yukun Liu[b], Moming Li[c], and Chiung-Yu Huang[a]

[a]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD; [b]KLATASDS-MOE and School of Statistics, East China Normal University, Shanghai, China; [c]Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, CA

ABSTRACT

Owing to its appealing distribution-free feature, conformal inference has become a popular tool for constructing prediction intervals with a desired coverage rate. In scenarios involving covariate shift, where the shift function needs to be estimated from data, many existing methods resort to data-splitting techniques. However, these approaches often lead to wider intervals and less reliable coverage rates, especially when dealing with finite sample sizes. To address these challenges, we propose methods based on a pivotal quantity derived under a parametric working model and employ a resampling-based framework to approximate its distribution. The resampling-based approach can produce prediction intervals with a desired coverage rate without splitting the data and can be easily applied to causal inference settings where a shift in the covariate distribution can occur between treatment and control arms. Additionally, the proposed approaches enjoy a double robustness property and are adaptable to different prediction tasks. Our extensive numerical experiments demonstrate that, compared to existing methods, the proposed novel approaches can produce substantially shorter conformal prediction intervals with lower variability in the interval lengths while maintaining promising coverage rates and advantages in versatile usage. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received May 2023
Accepted May 2024

KEYWORDS
Causal inference; Conformal inference; Covariate shift; Doubly robust; Prediction interval

1. Introduction

Conformal inference, also referred to as conformal prediction, has drawn much attention due to its ability to quantify prediction uncertainty, an often overlooked aspect in traditional machine learning methods. Introduced and formalized by Vovk, Gammerman, and Shafer (2005), conformal inference offers finite-sample coverage guarantees for predictions without employing distributional assumptions. This allows pre-trained machine learning models, such as random forest (Breiman 2001) and gradient boosting (Friedman 2001), to be integrated into the framework. Notable works in this area include Lei et al. (2018), Tibshirani et al. (2019), Romano, Patterson, and Candès (2019), Lei and Candès (2021), Chernozhukov, Wüthrich, and Zhu (2021), Barber et al. (2021a), and Barber et al. (2021b), among others. More recent works on conformal inference in various applications can be found in Fannjiang et al. (2022), Park et al. (2022), Qiu, Dobriban, and Tchetgen Tchetgen (2023), Yang, Kuchibhotla, and Tchetgen Tchetgen (2024), Yin et al. (2024), Candès, Lei, and Ren (2023), and Jin and Candès (2023). Interested readers can find comprehensive tutorials in Shafer and Vovk (2008) and Angelopoulos and Bates (2022).

Suppose the training data $\mathcal{D}_1 = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ are iid copies of a random vector $(\mathbf{X}, Y)$. The goal of conformal inference is to predict the value of a future outcome Y_{n+1} corresponding to a covariate vector $\mathbf{X}_{n+1}$ with a coverage rate of $1 - a$. Specifically, the goal is to determine the upper and lower limits of a prediction interval, denoted by $L(\mathbf{X}_{n+1})$ and $R(\mathbf{X}_{n+1})$, so that $\Pr\{L(\mathbf{X}_{n+1}) \le Y_{n+1} \le R(\mathbf{X}_{n+1})\} = 1 - a$, where $\mathbf{X}_{n+1}$ is a random draw from $\mathbf{X}$. Note that the coverage probability is marginalized over both $\mathbf{X}_{n+1}$ and Y_{n+1}. This is in contrast to prediction in regression settings, where $L(\mathbf{X}_{n+1})$ and $R(\mathbf{X}_{n+1})$ are chosen to satisfy $\Pr\{L(\mathbf{X}_{n+1}) \le Y_{n+1} \le R(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}\} = 1 - a$, and the probability is evaluated by conditioning on $\mathbf{X}_{n+1}$. In this case, a correct regression model for the conditional distribution of Y given $\mathbf{X}$ is required to guarantee the conditional coverage rate. In contrast, conformal inference leverages order statistics and their stochastic properties, eliminating the need for distributional assumptions.

Conformal prediction requires exchangeability of data points, under which the ranks of random variables are uniform over all possible permutations. However, shifts in distribution can break this exchangeability and compromise the finite-sample coverage rate of prediction intervals. This article focuses on the covariate shift problem, where the marginal distribution of $\mathbf{X}$ can vary, but the conditional distribution of Y given $\mathbf{X}$ remains constant across datasets. This problem arises in causal inference in observational studies, where randomization of treatments or interventions is challenging or infeasible. As a result, the composition of subjects receiving the new

CONTACT Yukun Liu ✉ ykliu@sfs.ecnu.edu.cn KLATASDS-MOE and School of Statistics, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China.

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

treatment can significantly differ from those who received the control treatment, leading to a shift in the marginal covariate distribution.

This article presents three conformal prediction intervals that can accommodate shifts in the covariate distribution. The proposed methods rely on an asymptotically pivotal statistic based on the cumulative distribution function (CDF) rather than commonly used residual-based nonconformity scores. The first method (see Section 2.2) is an unsplit version of conditional conformal inference under weighted exchangeability (Tibshirani et al. 2019), where it replaces the nonconformity score with the CDF in the conditional probability calculation. The other two methods are both resampling-based approaches. The second method (Section 2.3) employs exponential tilting of the empirical CDF in training data to estimate the CDF in testing data. We then approximate the distribution of the pivotal statistic by simulating training and testing data from the two estimated CDFs, which allows us to determine lower and upper quantiles for future outcome predictions given covariates. The third method, similar to the second, is tailored for causal inference in observational studies and incorporates control arm information to improve estimation of the joint CDF of covariates and potential outcomes under treatment in both arms. The resampling-based methods offer great flexibility and can be easily adapted to a variety of prediction tasks. Importantly, all three methods possess the double robustness properties, ensuring their validity when either the outcome regression model or the propensity score model (or weight function in the covariate-shift model) is correctly specified.

2. Conformal Prediction Methods

2.1. A Brief Review of Existing Methods

Conformal inference typically begins with a nonconformity score that measures how well an observation conforms to the rest of the data. A commonly used nonconformity score is the absolute residual obtained from a fitted model $\widehat{\mu}(\mathbf{x})$ for $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. The idea is to test the null hypothesis that $Y_{n+1} = y$ and construct a valid p-value based on the empirical quantiles of the nonconformity scores. A prediction interval is then formed from the collection of y values not rejected by the hypothesis test.

Three main approaches have been adopted for constructing prediction intervals. The full conformal prediction (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) leverages the exchangeability of $(\mathbf{X}_1, Y_1)$, …, $(\mathbf{X}_{n+1}, Y_{n+1})$ under the null hypothesis to ensure that absolute residuals $R_{y,i} = |Y_i - \widehat{\mu}_y(\mathbf{X}_i)|$, $i = 1, \ldots, n$, and $R_{y,n+1} = |y - \widehat{\mu}_y(\mathbf{X}_{n+1})|$ share the same distribution. Here the fitted model $\widehat{\mu}_y(\mathbf{x})$ is obtained using an augmented dataset $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ and $(\mathbf{X}_{n+1}, Y_{n+1})$ with $Y_{n+1} = y$ to avoid overfitting. A prediction interval is obtained based on the rank of the absolute residual for Y_{n+1} under the null hypothesis. Clearly, full conformal prediction is computationally intensive as it requires repeating the regression algorithm many times. In practice, a grid search is usually performed on a set of pre-specified values to reduce computational burden.

The second approach, known as split conformal prediction (Papadopoulos et al. 2002; Lei et al. 2018), involves dividing the training data into two disjoint subsets: a training subset $\mathcal{D}_1$ and a calibration subset $\mathcal{D}_2$. A regression model $\widehat{\mu}(\cdot)$ is trained on $\mathcal{D}_1$, and quantiles of the absolute residuals are evaluated and ranked using $\mathcal{D}_2$. Specifically, assuming an equal split, let d be the $\lceil n(1-a)/2 \rceil$th smallest value among $R_i = |Y_i - \widehat{\mu}(\mathbf{X}_i)|$, where $i \in \mathcal{D}_2$. Then, a split conformal prediction interval is given by $\{y \in \mathbb{R} : |y - \widehat{\mu}(\mathbf{X}_{n+1})| \le d\}$.

The data-splitting strategy, while avoiding grid search, compromises prediction efficiency because smaller calibration and training folds can lead to highly variable nonconformity scores and poor model fit. To address this issue, cross-conformal prediction methods have been developed (Vovk 2015). The algorithms divide the training data into K disjoint subsets, using one subset as calibration set and the remaining subsets for training the model. Thus, the evaluation of nonconformity scores exploits the full training dataset, akin to K-fold cross-validation. The special case $K = n$ corresponds to the Jackknife prediction interval (Lei et al. 2018). Denote by $\widehat{\mu}_{-i}(\cdot)$ a fitted model using the first n observations but with the ith observation removed. Let d be the $\lceil n(1-a) \rceil$th smallest values among the leave-one-out nonconformity scores $R_i = |Y_i - \widehat{\mu}_{-i}(\mathbf{X}_i)|$, $i = 1, \ldots, n$. Then a jackknife prediction interval can be given by $\{y \in \mathbb{R} : |y - \widehat{\mu}(\mathbf{X}_{n+1})| \le d\}$, where $\widehat{\mu}(\cdot)$ is derived from the first n observations. Barber et al. (2021b) showed that the Jackknife procedure is asymptotically valid with a stable estimator. Furthermore, they introduced a Jackknife+ procedure that provides finite-sample coverage guarantees without relying on assumptions that could be invalid in practical applications. Finally, Kim, Xu, and Barber (2020) introduced the jackknife+-after-bootstrap algorithm by leveraging ensemble learning and bootstrap methods.

2.2. Conditional Prediction Under Covariate Shift

We consider the setting where the training data share a joint density function $f_0(\mathbf{x}, y)$ and the new data point $(\mathbf{X}_{n+1}, Y_{n+1})$ is independently drawn from another joint density function $f_1(\mathbf{x}, y)$ with $f_1(\mathbf{x}, y) = w(\mathbf{x}) f_0(\mathbf{x}, y)$. Here $w(\mathbf{x})$ is a known, nonnegative weight function. It can be verified that the conditional distribution of Y_{n+1} given $\mathbf{X}_{n+1}$ is the same as that of Y_1 given $\mathbf{X}_1$, but their marginal covariate distributions differ. Moreover, the special case $w(\mathbf{x}) \equiv 1$ implies exchangeability between training and testing data. For ease of discussion, we assume that Y is continuous to avoid ties in the conditional probability calculation.

It is easy to see that $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+1}, Y_{n+1})\}$ are weighted exchangeable in the sense of Tibshirani et al. (2019) with weights $w_i = w(\mathbf{X}_i) = 1$, $i = 1, \ldots, n$, and $w_{n+1} = w(\mathbf{X}_{n+1})$. Hence, the weighted conformal inference can be carried out by comparing the value of a weighted nonconformity score at a test point to the weighted empirical CDF of nonconformity scores. Interestingly, the weighted empirical CDF can be viewed as the result of applying the conditioning technique to eliminate the nuisance function $f_0(\mathbf{x}, y)$. To see this, let Ψ denote the event that the collection of realized values in the training and testing data are $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+1}, Y_{n+1})\}$ without knowing which one belongs to the testing data. Then, for any nonconformity score $S(\mathbf{x}, y)$, the CDF of $S(\mathbf{X}_{n+1}, Y_{n+1})$ conditional on Ψ is

$$\Pr\{S(\mathbf{X}_{n+1}, Y_{n+1}) \le c \mid \Psi\}$$

$$= \frac{\sum_{i=1}^{n+1} w(\mathbf{X}_i) I\{S(\mathbf{X}_i, Y_i) \le c\}}{\sum_{j=1}^{n+1} w(\mathbf{X}_j)}, \quad \forall c \in \mathbb{R}, \tag{1}$$

and is thus free of the nuisance function $f(\mathbf{x}, y)$. Intuitively, a conformal prediction set with a target coverage rate of $1 - a$ can be given by $\{y : L_n \le S(\mathbf{X}_{n+1}, y) \le R_n\}$, where L_n and R_n are the $(a/2)$th and $(1 - a/2)$th quantiles of the CDF defined in (1). The data-splitting method is often used to guarantee finite-sample coverage, with a training fold used to train a model and a calibration fold used to determine quantiles. In practice, the weight function $w(\mathbf{x})$ is usually unknown and must be estimated from pooled covariate data. When consistently estimated, Lei and Candès (2021) pointed out that the coverage guarantee of the prediction interval is only valid asymptotically. Candès, Lei, and Ren (2023) further provided a general theory to derive nonasymptotic bounds for the coverage, subsequently establishing double robustness results for weighted conformal inference.

In this section, we present a new approach for constructing prediction intervals under covariate shift. Our proposed algorithm has two key features. First, it does not require data splitting. Second, instead of using a residual-based nonconformity score, we advocate using the conditional CDF of Y given $\mathbf{X}$ as a replacement in conformal inference. This strategy allows us to construct asymptotically pivotal statistics based on CDF. Specifically, we fit a parametric model $F(y \mid \mathbf{x}; \theta)$ and set $S(\mathbf{x}, y) = F(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})$ with $\widehat{\boldsymbol{\theta}}$ being a consistent estimator based on the training data. We further parameterize $w(\mathbf{x})$ as $w(\mathbf{x}; \boldsymbol{\beta})$ and obtain a consistent estimator $\widehat{\boldsymbol{\beta}}$ using the pooled covariate data. Although $S(\mathbf{X}_{n+1}, Y_{n+1})$ involves the unknown Y_{n+1}, we can set it to either 0 or 1 (lower and upper bounds of a CDF) without significantly affecting the evaluation of (1) as long as n is sufficiently large. Our numerical studies show that setting $S(\mathbf{X}_{n+1}, Y_{n+1})$ to either 0 or 1 produces satisfactory coverage, with $S(\mathbf{X}_{n+1}, Y_{n+1}) = 0$ often resulting in shorter prediction intervals than $S(\mathbf{X}_{n+1}, Y_{n+1}) = 1$.

Next, we show in Theorem 1 that the proposed method possesses a double robustness property: it provides asymptotically guaranteed marginal coverage when either the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ or the model for the weight function $w(\mathbf{x}; \boldsymbol{\beta})$ is correctly specified. A detailed proof can be found in Section 1 of the supplemental materials. Here, we provide a heuristic argument. When $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified, the distribution of $U_i = F(Y_i \mid \mathbf{X}_i; \widehat{\boldsymbol{\theta}})$, $i = 1, \dots, n + 1$, can be reasonably approximated by the uniform distribution U on $[0, 1]$. We set L_n and R_n to be the $(a/2)$th and $(1 - a/2)$th weighted quantiles of $U_i, i = 1, \dots, n$. Note that the last term $w(\mathbf{X}_{n+1}) = w(\mathbf{X}_{n+1}; \widehat{\boldsymbol{\beta}})$ can be ignored in the evaluation of (1) without affecting the large-sample result, and thus, approximately,

$$\frac{\sum_{i=1}^{n} w(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) I(U_i \le L_n)}{\sum_{j=1}^{n} w(\mathbf{X}_j; \widehat{\boldsymbol{\beta}})} = \frac{a}{2} \quad \text{and}$$

$$\frac{\sum_{i=1}^{n} w(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) I(U_i \le R_n)}{\sum_{j=1}^{n} w(\mathbf{X}_j; \widehat{\boldsymbol{\beta}})} = 1 - \frac{a}{2}.$$

It can be shown that the weighted empirical distribution of U_i's converges to $\Pr\{F(Y \mid \mathbf{X}; \boldsymbol{\theta}_*) \le t\}$ uniformly in $t \in [a_0, 1 - a_0]$ for any $a_0 \in (0, a/2)$, where $\boldsymbol{\theta}_*$ is the limit of $\widehat{\boldsymbol{\theta}}$ in probability. As $F(Y \mid \mathbf{X}; \boldsymbol{\theta}_*)$ is continuous, L_n and R_n must converge to some limits, denote by L and R, respectively. Combining the result that

$$\sum_{i=1}^{n} w(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}) I(U_i \le L_n) / \sum_{j=1}^{n} w(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}) \\ \to \Pr\{F(Y \mid \mathbf{X}; \boldsymbol{\theta}_*) \le L\} = a/2$$

and with the fact that $F(Y \mid \mathbf{X}; \boldsymbol{\theta}_*)$ follows the standard uniform distribution conditional on $\mathbf{X}$, we have $L = a/2$. Similarly, we can show that $R = 1 - a/2$. Therefore,

$$\Pr\{L_n \le F(Y_{n+1} \mid \mathbf{X}_{n+1}; \widehat{\boldsymbol{\theta}}) \le R_n\} \\ \to \Pr(a/2 \le U \le 1 - a/2) = 1 - a.$$

On the other hand, when the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is misspecified but the weight function $w(\mathbf{x}; \boldsymbol{\beta})$ is correctly specified, the conditional CDF in (1) is valid when $w(\mathbf{x})$ is replaced by $w(\mathbf{x}; \boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is the truth. As a result, it holds approximately when $w(\mathbf{x})$ is replaced by $w(\mathbf{x}; \boldsymbol{\beta})$ and a consistent estimate $\widehat{\boldsymbol{\beta}}$ is used. Thus, the proposed conformal prediction interval always has an approximately correct coverage, even when the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is misspecified.

Denote by P_0 the probability measure induced by $F_0(\mathbf{x}, y)$ and define $P_0 g(\mathbf{X}, Y) = \int g(\mathbf{x}, y) dF_0(\mathbf{x}, y)$ for any deterministic or random function g. Denote by $\mathcal{X}$, $\mathcal{Y}$, and Θ the parameter spaces of $\mathbf{X}$, Y and $\boldsymbol{\theta}$, respectively. We impose the following two sets of regularity conditions on the working models $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ and $w(\mathbf{x}; \boldsymbol{\beta})$.

(A) (i) Θ is a compact set in a Euclidean space; (ii) $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and has a probability density function $f(y \mid \mathbf{x}; \boldsymbol{\theta})$ (with respect to the Lebesgue or counting measure) that is continuous in $\boldsymbol{\theta}$ for each $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$; (iii) There exists a positive function $K(\mathbf{x}, y)$ such that $\sup_{\boldsymbol{\theta} \in \Theta} |\log f(y \mid \mathbf{x}; \boldsymbol{\theta})| \le K(\mathbf{x}, y)$ for each $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ and $P_0\{K(\mathbf{X}, Y)\} < \infty$.

(W1) (i) $w(\mathbf{x}; \boldsymbol{\beta})$ is continuous at $\boldsymbol{\beta}$ for each $\mathbf{x}$; (ii) The range $\mathcal{B}$ of $\boldsymbol{\beta}$ is a compact set in a Euclidean space; (iii) there exists $K_1(\mathbf{X})$ such that $P_0 K_1(\mathbf{X}) < \infty$ and $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \{w(\mathbf{x}; \boldsymbol{\beta})\}^2 \le K_1(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$

Condition (A) ensures that the function class $\{\log f(y \mid \mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is Glivenko–Cantelli (GC); see Example 19.8 of van der Vaart (1998). Condition (W1) implies that the function class $\{w(\mathbf{x}; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}$ is a GC class with a square integrable envelope and $n^{-1} \sum_{i=1}^{n} \{w(\mathbf{X}_i; \widehat{\boldsymbol{\beta}})\}^k = P_0\{w(\mathbf{X}; \widehat{\boldsymbol{\beta}})\}^k + o_p(1)$ for $k = 1, 2$. These results play a crucial role in the proof of Theorem 1, as elaborated in Section 1 of the supplementary materials.

Theorem 1. Assume that Conditions (A) and (W1) hold. The proposed conformal prediction set has an asymptotically correct coverage, that is, $\lim_{n \to \infty} \Pr\{L_n < S(\mathbf{X}_{n+1}, Y_{n+1}) \le R_n\} \ge 1 - a$, when one of the following two sets of conditions holds: (a) the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified with $\widehat{\boldsymbol{\theta}}$ being a consistent estimator for the unique true parameter value $\boldsymbol{\theta}_*$; (b) the weight function $w(\mathbf{x}; \boldsymbol{\beta})$ is correctly specified with $\widehat{\boldsymbol{\beta}}$ being a consistent estimator for the unique true parameter value $\boldsymbol{\beta}_*$.

2.3. Resampling-Based Prediction Under Covariate Shift

In the previous section, we assumed that the weight function $w(\mathbf{x})$ was either given or can be consistently estimated, without discussing the details of how to estimate it or whether the asymptotic properties can be affected by weight function parameter estimation. In this section, we propose a resampling-based procedure for conformal inference that incorporates the estimation of the weight function and does not require data splitting. Specifically, we assume that the joint density functions of the training and testing data, that is, $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y)$, follow an exponential tilt model

$$\frac{f_1(\mathbf{x}, y)}{f_0(\mathbf{x}, y)} = \exp\{\alpha + \boldsymbol{\beta}^\top @@ \boldsymbol{\phi}(\mathbf{x})\}, \tag{2}$$

where $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^q$ is a pre-specified q-dimensional vector function, $\boldsymbol{\beta}$ is a q-dimensional parameter, and α satisfies $\iint f_0(\mathbf{x}, y) \exp\{\alpha + \boldsymbol{\beta}^\top \boldsymbol{\phi}(\mathbf{x})\} d\mathbf{x} dy = 1$ to ensure that $f_1(\mathbf{x}, y)$ is a proper density function. The proposed model is equivalent to setting $w(\mathbf{x}) = \exp\{\alpha + \boldsymbol{\beta}^\top \boldsymbol{\phi}(\mathbf{x})\}$ in the covariate shift model considered in the previous section. For ease of exposition, we set $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$ with the understanding that the results established in this section can be easily extended to a more general $\boldsymbol{\phi}$. Under (2), the training and testing data share the same conditional distribution function of the outcome given covariates, denoted by $F(y \mid \mathbf{x})$, while the marginal distribution of the covariate is allowed to differ.

As mentioned before, one can construct asymptotically pivotal statistics based on CDF. To see this, suppose the distribution of Y given $\mathbf{X} = \mathbf{x}$ is known up to a finite-dimensional parameter $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ were known, $U(\boldsymbol{\theta}) = F(Y \mid \mathbf{X}; \boldsymbol{\theta})$ follows a uniform distribution on $[0, 1]$. Then, for $a \in (0, 1)$, a $100(1-a)\%$ prediction interval for Y_{n+1} is given by $\{y : a/2 \le F(y \mid \mathbf{X}_{n+1}; \boldsymbol{\theta}) \le 1 - a/2\}$. In practice, $\boldsymbol{\theta}$ is unknown and needs to be estimated using the training data. Let $\widehat{\boldsymbol{\theta}}$ be an estimator for $\boldsymbol{\theta}$. When $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified, then $U(\widehat{\boldsymbol{\theta}})$ is approximately uniformly distributed on $[0, 1]$ when n is large. As pointed out in Cox (1975), however, the uniform distribution may not approximate the distribution of $U(\widehat{\boldsymbol{\theta}})$ well with a small n. To see this, let $G(u)$ be the CDF of $U(\widehat{\boldsymbol{\theta}})$ so that

$$\begin{aligned} G(u) &= \Pr\{U(\widehat{\boldsymbol{\theta}}) \le u\} \\ &= \Pr\{F(Y \mid \mathbf{X}; \widehat{\boldsymbol{\theta}}) \le u\} \\ &= \Pr\{Y \le F^{-1}(u \mid \mathbf{X}; \widehat{\boldsymbol{\theta}})\}, \end{aligned}$$

where the plug-in estimator $\widehat{\boldsymbol{\theta}}$ is treated as a random variable. Define $\widetilde{F}(y \mid \mathbf{x}) = G\{F(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})\}$ and its corresponding density function $\widetilde{f_p}(y \mid \mathbf{x}) = g\{F(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})\} f(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}}) = g\{U(\widehat{\boldsymbol{\theta}})\} f(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})$. When U is exactly pivotal, that is, the distribution of U is independent of $\boldsymbol{\theta}$, Harris (1989) showed that $\widetilde{f_p}(y \mid \mathbf{x})$ dominates the plug-in density $f(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})$ in terms of average Kullback–Leibler distance. When n is in the range of 10–50, simulation results reported in Lawless and Fredette (2005) indicated that the prediction intervals derived from $\widetilde{F}(y \mid \mathbf{x})$ have better coverage than those derived from the plug-in function $F(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})$ that ignores the randomness in $\widehat{\boldsymbol{\theta}}$. On the other hand, if n is moderate/large, these two methods give almost identical results. Recently, Tian et al. (2022) proposed a calibration-bootstrap procedure by repeatedly sampling from $F(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})$ to calibrate the plug-in prediction intervals. However, their method is valid only under the exchangeability assumption between training and testing data.

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be the training data and $(\mathbf{X}_{n+1}, Y_{n+1}), \ldots, (\mathbf{X}_N, Y_N)$ be the testing data, where outcome values Y_i, $i = n+1, \ldots, N$, in the testing data are not available. Define $F_0(\mathbf{x}, y)$ and $F_1(\mathbf{x}, y)$ the cumulative distribution functions corresponding to $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y)$, respectively. Denote by p_i the jump size of $F_0(\mathbf{x}, y)$ at $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$. In the ideal situation where the outcomes in the testing data were available, the log-likelihood function based on the pooled data $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, N\}$ is

$$\sum_{i=1}^{n} \log p_i + \sum_{i=n+1}^{N} \left(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i + \log p_i\right),$$

where p_i's satisfy the constraints $p_i \ge 0$, $\sum_{i=1}^N p_i = 1$, and $\sum_{i=1}^N p_i \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i) = 1$ to ensure that $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y) = f_0(\mathbf{x}, y) \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{x})$ are proper density functions. Profiling out p_i subject to these constraints yields (see, e.g., Qin 2017, sec. 11.1)

$$p_i = \frac{1}{n} \times \frac{1}{1 + \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i) \cdot (N-n)/n}, \quad i = 1, \ldots, N,$$

and the log profile likelihood, up to a constant,

$$\begin{aligned} \widetilde{\ell}(\alpha, \boldsymbol{\beta}) = &-\sum_{i=1}^{N} \log\left\{1 + \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i) \cdot (N-n)/n\right\} \\ &+ \sum_{i=n+1}^{N} \left(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i\right). \end{aligned} \tag{3}$$

We denote the maximizer of $\widetilde{\ell}(\alpha, \boldsymbol{\beta})$ by $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$, and denote

$$\widehat{p}_i = \frac{1}{n} \times \frac{1}{1 + \exp(\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i) \cdot (N-n)/n}, \quad i = 1, \ldots, N.$$

Note that the profile likelihood (3) does not require any knowledge of the outcomes, thus, the parameters $(\alpha, \boldsymbol{\beta})$ in the covariate-shift model (2) can be estimated by $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ without knowing Y_i's. Moreover, if the values of $Y_{n+1}, \ldots, Y_N$ were available, one could estimate F_0 and F_1 by $\widetilde{F}_0(\mathbf{x}, y) = \sum_{i=1}^N \widehat{p}_i I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)$ and $\widetilde{F}_1(\mathbf{x}, y) = \sum_{i=1}^N \widehat{p}_i \exp(\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i) I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)$, respectively, where $\le$ is applied componentwise for a vector.

To address the challenge that Y_i's in the testing data are not available for evaluating $\widetilde{F}_1(\mathbf{x}, y)$, we propose to use the covariate-shift model (2) to construct a consistent estimator. By leveraging the fact that $F_1(\mathbf{x}, y) = \iint_{\mathbf{u} \le \mathbf{x}, v \le y} \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{u}) dF_0(d\mathbf{u}, dv)$ under model (2), we can estimate $F_1(\mathbf{x}, y)$ as

$$\begin{aligned} &\iint_{\mathbf{u} \le \mathbf{x}, v \le y} \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{x}) d\widehat{F}_0(d\mathbf{u}, dv) \\ &\quad = n^{-1} \sum_{i=1}^{n} \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i) I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y), \end{aligned}$$

where $\widehat{F}_0(\mathbf{x}, y) = n^{-1} \sum_{i=1}^n I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)$ is the empirical CDF. To ensure the resulting estimator yields a proper distribution function, we consider normalized weights $\exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_i) / \sum_{k=1}^n \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}_k) = \exp(\boldsymbol{\beta}^\top \mathbf{X}_i) / \sum_{k=1}^n \exp(\boldsymbol{\beta}^\top \mathbf{X}_k)$ and obtain

$$\widehat{F}_1(\mathbf{x}, y) = \frac{\sum_{i=1}^n \exp(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i) I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)}{\sum_{j=1}^n \exp(\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_j)}$$

Algorithm 1: Resampling-based conformal prediction set in the presence of covariate shift

```
Data: ${(\mathbf{X}_i, Y_i) : 1, \ldots, n}$ and ${\mathbf{X}_j : j = n+1, \ldots, N}$
Input: $B$: number of bootstrap samples; $1-a$: the target coverage rate; a parametric working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$; a
        covariate-shift model (2)
Output: Prediction sets for $Y_j, j = n+1, \ldots, N$
Prepare
  • Obtain $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ by maximizing (3) under the model (2) based on ${\mathbf{X}_i : i = 1, \ldots, N}$
  • Obtain the MLE $\widehat{\boldsymbol{\theta}}$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on ${(\mathbf{X}_i, Y_i) : i = 1, \ldots, n}$
  • Obtain the empirical CDF $\widehat{F}_0(\mathbf{x}, y) = n^{-1} \sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)$ and $\widehat{F}_1(\mathbf{x}, y)$ defined in (4) under the model (2)

for $b = 1 : B$ do
    • Sample ${(\mathbf{X}_i^b, Y_i^b) : i = 1, \ldots, n}$ from $\widehat{F}_0(\mathbf{x}, y)$ and ${(\mathbf{X}_j^b, Y_j^b) : j = n+1, \ldots, N}$ from $\widehat{F}_1(\mathbf{x}, y)$
    • Obtain the MLE $\widehat{\boldsymbol{\theta}}^b$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on ${(\mathbf{X}_i^b, Y_i^b) : i = 1, \ldots, n}$
    • Calculate $U_j^b = F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b), j = n+1, \ldots, N$

For $j = n+1, \ldots, N$, obtain $L_{nj}$ and $R_{nj}$, the $\lfloor Ba/2 \rfloor$th and $\lceil B(1-a/2) \rceil$th smallest values of ${U_j^b : b = 1, \ldots, B}$,
respectively
Result: A conformal prediction set for $Y_j$ with a coverage rate of $1-a$ is given by ${y : L_{nj} \le F(y \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) \le R_{nj}}$
```

$$= \frac{\iint_{\mathbf{u} \le \mathbf{x}, v \le y} \exp(\widehat{\boldsymbol{\beta}}^\top \mathbf{u}) d\widehat{F}_0(\mathbf{u}, v)}{\iint \exp(\widehat{\boldsymbol{\beta}}^\top \mathbf{s}) d\widehat{F}_0(\mathbf{s}, t)}. \tag{4}$$

When the covariate-shift model (2) is correctly specified, the resulting semiparametric estimator for $F_1(\mathbf{x}, y)$ is uniformly consistent and asymptotically normal, as summarized in Theorem 2. Throughout the article, we make the assumption that $n/N = \rho + o(n^{-1/2})$ for a constant $\rho \in (0, 1)$.

Theorem 2. Assume that the range $\mathcal{X} \times \mathcal{Y}$ of $(\mathbf{X}, Y)$ is compact and that the covariate-shift model (2) is correctly specified, whose true parameter values are denoted by $(\alpha_0, \boldsymbol{\beta}_0)$. Suppose that $\sup_{\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}_0, \delta_0)} P_0 \exp(\boldsymbol{\beta}^\top \mathbf{X}) < \infty$ for some $\delta_0 > 0$, where $\mathcal{N}(\boldsymbol{\beta}_0, \delta) = \{\boldsymbol{\beta} :\mid \boldsymbol{\beta} - \boldsymbol{\beta}_0 \mid \le \delta\}$, and that the matrix $\boldsymbol{A}_1 = P_0[(1, \mathbf{X}^\top)^\top (1, \mathbf{X}^\top) \rho \exp(\alpha_0 + \mathbf{X}^\top \beta_0) / \{\rho + (1-\rho) \exp(\alpha_0 + \mathbf{X}^\top \beta_0)\}]$ is positive definite. Then, as $n \to \infty$, the following results hold: (a) $\sqrt{N}\{(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})^\top - (\alpha_0, \boldsymbol{\beta}_0)^\top\} \xrightarrow{d} N(0, \Omega)$, where $\Omega = \boldsymbol{A}_1^{-1} - \rho^{-1} \mathbf{e}_1 \mathbf{e}_1^\top$ and $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top$ is a $(p+1)$-dimensional vector. (b) $\widehat{F}_1(\mathbf{x}, y)$ converges uniformly to $F_1(\mathbf{x}, y)$ in probability, that is, $\sup_{(\mathbf{x}, y)} |\widehat{F}_1(\mathbf{x}, y) - F_1(\mathbf{x}, y)| = o_p(1)$. (c) The stochastic process $\{\sqrt{n}\{\widehat{F}_1(\mathbf{x}, y) - F_1(\mathbf{x}, y)\} : (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}\}$ converges weakly to a mean zero Gaussian process.

Let $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ be a working model for the conditional CDF $F(y \mid \mathbf{x})$, and denote by $\widehat{\boldsymbol{\theta}}$ the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$ obtained using the training data. Define $F^{-1}(t \mid \mathbf{x}; \boldsymbol{\theta}) = \inf\{y \in \mathbb{R} : F(y \mid \mathbf{x}; \boldsymbol{\theta}) \ge t\}$. Our resampling-based conformal prediction sets for $Y_j, j = n+1, \ldots, N$, can be constructed using Algorithm 1.

The proposed algorithm enjoys a double robustness property: the prediction sets have an asymptotically correct coverage rate if either the working regression model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ or the covariate-shift model (2) is correctly specified. To see this, Theorem 2 implies that $\widehat{F}_1(\mathbf{x}, y)$ is a consistent estimator for $F_0(\mathbf{x}, y)$ when the covariate-shift model is correctly specified. As a result, the resampled data $\{(\mathbf{X}_j^b, Y_j^b) : j = n+1, \ldots, N\}$ approximately follows the same distribution as $\{(\mathbf{X}_j, Y_j) : j = n+1, \ldots, N\}$ in the testing data. Thus, $U_j^b = F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b)$ and $U_j = F(Y_j \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}})$ approximately share the same distribution, regardless of whether the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified. Therefore, the prediction sets given by Algorithm 1 have an approximately correct coverage rate.

On the other hand, when the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified, it is easy to see that both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^b$ are consistent estimates of the true parameter value $\boldsymbol{\theta}_0$. Denote by $(\alpha_*, \boldsymbol{\beta}_*)$ the maximizer of $\ell_*(\alpha, \boldsymbol{\beta}) = P_0[(1-\rho)(\alpha + \boldsymbol{\beta}^\top \mathbf{X}) - \log\{1 + \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{X}) \times (1-\rho)/\rho\}]$, and define $F_1^*(\mathbf{x}, y) = P_0\{\exp(\boldsymbol{\beta}_*^\top \mathbf{X}) I(\mathbf{X} \le \mathbf{x}, Y \le y)\} / P_0\{\exp(\boldsymbol{\beta}_*^\top \mathbf{X})\}$. Under the misspecified covariate-shift model, $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}_*$ and that $\widehat{F}_1(\mathbf{x}, y)$ converges in probability to $F_1^*(\mathbf{x}, y)$ with joint density function $f_1^*(\mathbf{x}, y) = \exp(\boldsymbol{\beta}_*^\top \mathbf{x}) f_0(\mathbf{x}, y) / P_0\{\exp(\boldsymbol{\beta}_*^\top \mathbf{X})\}$ and marginal density function $g_1^*(\mathbf{x}) = \exp(\boldsymbol{\beta}_*^\top \mathbf{x}) \int f_0(\mathbf{x}, y) dy / P_0\{\exp(\boldsymbol{\beta}_*^\top \mathbf{X})\}$. Thus, the conditional density function is $f_1^*(y \mid \mathbf{x}) = f_1^*(\mathbf{x}, y) / g_1^*(\mathbf{x}) = f_0(y \mid \mathbf{x}) = f(y \mid \mathbf{x})$. Intuitively, the conditional distribution function of Y_j^b given $\mathbf{X}_j^b$ is approximately $F(y \mid \mathbf{x}_j^b)$, and hence $F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b)$ has an approximately uniform distribution on $[0, 1]$. As a result, for a fixed j, $\{U_j^b = F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b) : b = 1, \ldots, B\}$ are approximately iid and approximately follow the uniform distribution. Moreover, the distribution of $F(Y_j \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}})$ is also approximately uniform on $[0, 1]$ when the working model is correctly specified. Therefore, we can prove that the proposed prediction set has an approximately correct coverage rate even when the covariate-shift model is misspecified. Theorem 3 summarizes the asymptotic results and the desired doubly robustness property.

Theorem 3. Assume that Conditions (A) and (W1) hold. The prediction sets given by Algorithm 1 have an asymptotically

Algorithm 2: Resampling-based conformal prediction set in the presence of general covariate shift

```
Data: ${(\mathbf{X}_i, Y_i) : 1, \dots, n}$ and ${\mathbf{X}_j : j = n+1, \dots, N}$
Input: $B$: number of bootstrap samples; $1 - a$: the target coverage rate; a parametric working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$; an
       estimation procedure $\mathcal{A}$ for $w(\mathbf{x})$
Output: Prediction sets for $Y_j, j = n+1, \dots, N$
Prepare
  • Obtain $\widehat{w}(\mathbf{x})$ using the estimation procedure $\mathcal{A}$ based on ${\mathbf{X}_i : i = 1, \dots, N}$
  • Obtain the MLE $\widehat{\boldsymbol{\theta}}$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on ${(\mathbf{X}_i, Y_i) : i = 1, \dots, n}$
  • Obtain the empirical CDF $\widehat{F}_0(\mathbf{x}, y) = n^{-1} \sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)$ and $\widehat{F}_1(\mathbf{x}, y)$ defined in (5)

for $b = 1 : B$ do
    • Sample ${(\mathbf{X}_i^b, Y_i^b) : i = 1, \dots, n}$ from $\widehat{F}_0(\mathbf{x}, y)$ and ${(\mathbf{X}_j^b, Y_j^b) : j = n+1, \dots, N}$ from $\widehat{F}_1(\mathbf{x}, y)$
    • Obtain the MLE $\widehat{\boldsymbol{\theta}}^b$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on ${(\mathbf{X}_i^b, Y_i^b) : i = 1, \dots, n}$
    • Calculate $U_j^b = F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b), j = n+1, \dots, N$

For $j = n+1, \dots, N$, obtain $L_{nj}$ and $R_{nj}$, the $\lfloor Ba/2 \rfloor$th and $\lceil B(1 - a/2) \rceil$th smallest values of ${U_j^b : b = 1, \dots, B}$,
respectively
Result: A conformal prediction set for $Y_j$ with coverage rate $1 - a$ is given by ${y : F^{-1}(L_{nj} \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) \le y \le F^{-1}(R_{nj} \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}})}$
```

correct coverage rate either when the working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ for the conditional CDF is correctly specified with $\boldsymbol{\theta}_*$ being the unique true value of $\boldsymbol{\theta}$, or when the weight function $w(\mathbf{x}; \boldsymbol{\beta}) = \exp(\alpha + \boldsymbol{\beta}^\top \mathbf{x})$ is correctly specified with $(\alpha_*, \boldsymbol{\beta}_*)$ being the unique true value of $(\alpha, \boldsymbol{\beta})$.

To avoid misspecification of the covariate-shift model (2), one can employ modern machine learning methods such as artificial neural networks, random forests, and kernel methods to obtain a consistent estimate of the covariate-shift function $w(\mathbf{x})$. Given any estimator, say $\widehat{w}(\mathbf{x})$, of $w(\mathbf{x})$, we estimate $F_1(\mathbf{x}, y)$ by

$$\widehat{F}_1(\mathbf{x}, y) = \frac{\sum_{i=1}^{n} \widehat{w}(\mathbf{X}_i) I(\mathbf{X}_i \le \mathbf{x}, Y_i \le y)}{\sum_{j=1}^{n} \widehat{w}(\mathbf{X}_j)} = \frac{\iint_{\mathbf{u} \le \mathbf{x}, v \le y} \widehat{w}(\mathbf{u}) d\widehat{F}_0(\mathbf{u}, v)}{\iint \widehat{w}(\mathbf{s}) d\widehat{F}_0(\mathbf{s}, t)}. \tag{5}$$

We impose the following conditions on $\widehat{w}(\mathbf{x})$.

(W2) (i) $P_0 \|\widehat{w}(\cdot) - w(\cdot)\|^2 = o_p(1)$. (ii) $\mathscr{W}$ is a GC class of functions such that $\widehat{w}(\cdot) \in \mathscr{W}$. (iii) There exists a function $K_2(\mathbf{x})$ such that $\widetilde{w}(\mathbf{x}) \le K_2(\mathbf{x}), \mathbf{x} \in \mathcal{X}$ for all $\widetilde{w}(\cdot) \in \mathscr{W}$ and $P_0 K_2(\mathbf{X}) < \infty$.

Theorem 4. Assume that Conditions (A) and (W2) hold. The prediction sets given by Algorithm 2 always have an asymptotically correct coverage rate, irrespective of the correctness of the outcome regression model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$.

The proposed algorithm is highly flexible and applicable to a wide range of problems. For illustration, consider the task of predicting the maximum of the next k outcome values in the testing data, denoted by $Z = \max\{Y_{n+1}, \dots, Y_{n+k}\}$. One can leverage the fact that

$$\Pr(Z \le t \mid \mathbf{X}_{n+1}, \dots, \mathbf{X}_{n+k}) = \prod_{j=1}^{k} F(t \mid \mathbf{X}_{n+j}),$$

and replace the nonconformity score $F(Y_{n+j} \mid \mathbf{X}_{n+j}; \widehat{\boldsymbol{\theta}})$ in Algorithms 1 and 2 with $\prod_{j=1}^{k} F(Z \mid \mathbf{X}_{n+j}; \widehat{\boldsymbol{\theta}})$ to construct a conformal prediction set for Z. This demonstrates the flexibility and generalizability of the proposed procedure, which can be easily adapted to different prediction tasks by modifying the target variable and its corresponding probability distribution function. Additionally, the corresponding conformal prediction set for Z has a desirable double robustness property, similar to the proposed intervals for individual outcomes. This property is formally stated in the following corollary.

Corollary 1. (1) The proposed prediction set for Z maintains an asymptotically correct coverage rate under Conditions (A) and (W1), if either $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified with $\boldsymbol{\theta}_*$ being the unique true value of $\boldsymbol{\theta}$ or model (2) is correctly specified with $(\alpha_*, \boldsymbol{\beta}_*)$ being the unique true value of $(\alpha, \boldsymbol{\beta})$. (2) If we do not assume model (2) and take $\widehat{F}_1(\mathbf{x}, y)$ to be (5) instead of (4) in the proposed algorithm for the maximum of several responses, the resulting prediction set always has an asymptotically correct coverage rate under Conditions (A) and (W2), irrespective of the correctness of the outcome regression model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$.

3. Causal Inference With Conformal Prediction

In this section, we explore the application of conformal prediction to causal inference in observational studies. It is worth noting that the training and testing data in Section 2 are samples of fixed sizes, while sample sizes in the treatment and control groups are usually random in observational studies. As a result, there is a subtle difference in establishing the large sample properties of the estimated CDFs. Additionally, the empirical CDF estimated in Section 2.3 was based only on data from the training data, thus, can be inefficient. In this section, we consider a more efficient estimator that uses data from both the control and treatment groups, and investigate whether improving the efficiency of the CDF estimator can enhance the performance of the conformal prediction sets.

We adopt the potential outcome framework to describe data from observational studies. Let $Y(1)$ and $Y(0)$ denote the potential outcomes had an individual received active and control treatment, respectively. The treatment indicator D takes the value of 1 if the individual received active treatment and $D = 0$ otherwise. Consequently, we observe $Y = DY(1) + (1-D)Y(0)$ instead of completely observing both $Y(0)$ and $Y(1)$ at the same time. Note that conventional causal inference focuses on the marginal difference between treatment and control. In this work, we focus on predicting the unobserved potential outcome $Y(d)$ for individuals who received treatment $1-d$, $d = 0, 1$.

Denote by $\mathbf{X}$ a p-dimensional vector of covariates that may be correlated with the treatment received and the potential outcomes. Suppose $\{(Y_i(0), Y_i(1), \mathbf{X}_i, D_i) : i = 1, \dots, N\}$ are N iid copies of $(Y(0), Y(1), \mathbf{X}, D)$. Without loss of generality, we assume that the first $n = \sum_{i=1}^N D_i$ individuals received an active treatment (the treatment arm) and the remaining $N - n$ individuals received standard care (the control arm). For ease of exposition, we focus on the prediction of the unobserved potential outcome $Y(1)$, in the control arm based on data from the treatment group $\{(Y_i(1), \mathbf{X}_i, D_i = 1) : i = 1, \dots, n\}$ and the covariate data $\{(\mathbf{X}_i, D_i = 0) : i = n+1, \dots, N\}$ from the control group. A similar procedure can be used to predict $Y(0)$ in the treatment arm.

We adopt the standard assumptions from the causal inference literature:

(B) (i) (*Stable Unit Treatment Value*) The potential outcomes for any individual do not vary with the treatments assigned to other individuals. (ii) (*Unconfoundedness*) The treatment indicator D is conditionally independent of the potential outcomes $\{Y(0), Y(1)\}$ given $\mathbf{X}$. (iii) (*Overlap*) $0 < \Pr(D = 1 \mid \mathbf{X}) < 1$ for all $\mathbf{X}$.

Assuming unconfoundedness, the propensity score is given by $\Pr(D = 1 \mid Y(0), Y(1), \mathbf{X}) = \Pr(D = 1 \mid \mathbf{X}) := \pi(\mathbf{X})$ and a logistic model is commonly imposed for D given $\mathbf{X}$:

$$\pi(\mathbf{x}) = \pi(\mathbf{x}; \alpha, \boldsymbol{\beta}) := \frac{\exp(\alpha + \mathbf{x}^\top \boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x}^\top \boldsymbol{\beta})}, \qquad (6)$$

where $(\alpha, \boldsymbol{\beta})$ are unknown regression parameters that can be estimated by the MLE:

$$(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \underset{(\alpha,\boldsymbol{\beta})}{\operatorname{argmax}} \sum_{i=1}^N \Big[D_i \log\{\pi(\mathbf{X}_i; \alpha, \boldsymbol{\beta})\} + (1 - D_i) \log\{1 - \pi(\mathbf{X}_i; \alpha, \boldsymbol{\beta})\} \Big]. \qquad (7)$$

As discussed in the previous section, a crucial step in the resampling-based conformal prediction procedure is to estimate the joint CDF in the training data, which is $F_1(\mathbf{x}, y) = \Pr(\mathbf{X} \le \mathbf{x}, Y \le y \mid D = 1) = \Pr(\mathbf{X} \le \mathbf{x}, Y(1) \le y \mid D = 1)$ in the setting considered here. However, as pointed out in Qin (2017) (p. 211, Remark 3), the empirical CDF $\widehat{F}_1(\mathbf{x}, y) = n^{-1} \sum_{i=1}^n I(\mathbf{X}_i \le \mathbf{x}, Y_i(1) \le y)$, though a consistent estimator of F_1, may be inefficient under the unconfoundedness assumption because it fails to leverage information in the control group covariate data $\{\mathbf{X}_{n+1}, \dots, \mathbf{X}_N\}$. By employing the empirical likelihood method (Owen 1990), one can construct an improved estimator by incorporating such information. We will examine the impact of this improved estimator on prediction intervals.

We first note that the empirical CDF $F_1(\mathbf{x}, y)$ can be viewed as a maximum empirical likelihood estimator. Specifically, denoting p_i as the jump size of $F_1(\mathbf{x}, y)$ at $(\mathbf{X}_i, Y_i(1))$, the empirical CDF maximizes the nonparametric likelihood $\prod_{i=1}^n p_i$ constructed based on treatment group data with respect to the constraints $p_i \ge 0$ and $\sum_{i=1}^n p_i = 1$. Interestingly, under the unconfoundedness assumption, additional constraints can be incorporated to leverage information from the control group and improve efficiency (Qin, Liu, and Li 2023). Specifically, define the joint CDF $F(\mathbf{x}, y) = \Pr(\mathbf{X} \le \mathbf{x}, Y(1) \le y)$ and define $\Delta = \Pr(D = 1)$. Then the unconfoundedness assumption implies $dF_1(\mathbf{x}, y) = \{\pi(\mathbf{x})/\Delta\} dF(\mathbf{x}, y)$, leading to

$$\iint h(\mathbf{x}) dF_1(\mathbf{x}, y) = \frac{1}{\Delta} \iint h(\mathbf{x}) \pi(\mathbf{x}) dF(\mathbf{x}, y) \qquad (8)$$

for any $h(\mathbf{x})$ such that the above integral is bounded. A natural choice of h is $h(\mathbf{x}; \widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) = \exp(-\widehat{\alpha} - \mathbf{x}^\top \widehat{\boldsymbol{\beta}}) \mu(\mathbf{x}; \widehat{\boldsymbol{\theta}})$, where $\mu(\mathbf{x}; \boldsymbol{\theta})$ is a given working model for the conditional mean $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}, D = 1)$ and $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ obtained using the treatment group data. We can approximate the left-hand side of (8) by $n^{-1} \sum_{i=1}^n h(\mathbf{X}_i) p_i$ and the right-hand side by the empirical average $(n/N)^{-1} \times N^{-1} \sum_{i=1}^N h(\mathbf{X}_i) \pi(\mathbf{X}_i; \widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ over all available covariate data. To obtain the constrained MLE for $F_1(\mathbf{x}, y)$, we maximize the log empirical likelihood $\ell = \sum_{i=1}^n \log(p_i)$ subject to constraints $p_i \ge 0$, $\sum_{i=1}^n p_i = 1$, and $\sum_{i=1}^n p_i \boldsymbol{\psi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = 0$, where $\boldsymbol{\psi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\boldsymbol{\beta}}) = \mathbf{h}(\mathbf{X}_i) - n^{-1} \sum_{j=1}^N \mathbf{h}(\mathbf{X}_j) \pi(\mathbf{X}_j; \widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ and $\mathbf{h}$ can be vector-valued. The resulting estimator is (see, e.g., Qin 2017, chap. 19)

$$\widehat{F}_1(\mathbf{x}, y) = \sum_{i=1}^n \widehat{p}_i I(\mathbf{X}_i \le \mathbf{x}, Y_i(1) \le y), \qquad (9)$$

where $\widehat{p}_i = n^{-1}\{1 + \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{\psi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\boldsymbol{\beta}})\}^{-1}$ and $\widehat{\boldsymbol{\lambda}}$ satisfies $\sum_{i=1}^n \boldsymbol{\psi}(\mathbf{X}_i)/\{1 + \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{\psi}(\mathbf{X}_i; \widehat{\alpha}, \widehat{\boldsymbol{\beta}})\} = 0$. It is worth noting that the empirical likelihood framework allows for the incorporation of multiple working models simultaneously.

We next consider the estimation of $F_0(\mathbf{x}, y) = \Pr(\mathbf{X} \le \mathbf{x}, Y(1) \le y \mid D = 0)$. The logistic regression model (6) for the propensity score implies

$$\begin{aligned} dF_0(\mathbf{x}, y) &= \frac{1 - \pi(\mathbf{x})}{1 - \Delta} dF(\mathbf{x}, y) \\ &= \frac{1 - \pi(\mathbf{x})}{1 - \Delta} \times \frac{\Delta}{\pi(\mathbf{x})} dF_1(\mathbf{x}, y) \\ &= \frac{\exp(-\mathbf{x}^\top \boldsymbol{\beta}) dF_1(\mathbf{x}, y)}{\iint \exp(-\mathbf{s}^\top \boldsymbol{\beta}) dF_1(\mathbf{s}, t)}, \end{aligned}$$

which motivates the following estimator for $F_0(\mathbf{x}, y)$:

$$\widehat{F}_0(\mathbf{x}, y) = \frac{\sum_{i=1}^n \widehat{p}_i \exp(-\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}) I(\mathbf{X}_i \le \mathbf{x}, Y_i(1) \le y)}{\sum_{j=1}^n \widehat{p}_j \exp(-\mathbf{X}_j^\top \widehat{\boldsymbol{\beta}})}. \qquad (10)$$

The asymptotic properties of $\widehat{F}_1$ and $\widehat{F}_0$ are summarized in the following theorem.

Algorithm 3: Resampling-based conformal prediction set for causal inference

```
Data: $\{(Y_i(1), \mathbf{X}_i, D_i = 1) : 1, \ldots, n\}$ and $\{(\mathbf{X}_j, D_j = 0) : j = n+1, \ldots, N\}$
Input: $B$: number of bootstrap samples; $1 - a$: the target coverage rate; a parametric working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$; a propensity
        score model in (6)
Output: Prediction sets for $Y_j(1), j = n+1, \ldots, N$
Prepare
  • Obtain $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ defined in (7) based on $\{(\mathbf{X}_i, D_i) : i = 1, \ldots, N\}$
  • Obtain the MLE $\widehat{\boldsymbol{\theta}}$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$
  • Obtain $\widehat{F}_1(\mathbf{x}, y)$ and $\widehat{F}_0(\mathbf{x}, y)$ defined in (9) and (10), respectively

for $b = 1 : B$ do
    • Sample $\{(\mathbf{X}_i^b, Y_i^b) : i = 1, \ldots, n\}$ from $\widehat{F}_1(\mathbf{x}, y)$ and $\{(\mathbf{X}_j^b, Y_j^b) : j = n+1, \ldots, N\}$ from $\widehat{F}_0(\mathbf{x}, y)$
    • Obtain the MLE $\widehat{\boldsymbol{\theta}}^b$ under the model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ based on $\{(\mathbf{X}_i^b, Y_i^b) : i = 1, \ldots, n\}$
    • Calculate $U_j^b = F(Y_j^b \mid \mathbf{X}_j^b; \widehat{\boldsymbol{\theta}}^b), j = n+1, \ldots, N$

For $j = n+1, \ldots, N$, obtain $L_{nj}$ and $R_{nj}$, the $\lfloor Ba/2 \rfloor$th and $\lceil B(1 - a/2) \rceil$th smallest values of $\{U_j^b : b = 1, \ldots, B\}$,
respectively
Result: A conformal prediction set for $Y_j(1)$, of which $D_j = 0$, with a coverage rate of $1 - a$ is given by
        $\{y : L_{nj} \le F(y \mid \mathbf{X}_j; \widehat{\boldsymbol{\theta}}) \le R_{nj}\}$.
```

Theorem 5. Assume that Condition (B) holds, the propensity score model (6) is correctly specified with the true parameter values $(\alpha_0, \boldsymbol{\beta}_0)$ and that $\Delta = \Pr(D = 1) > 0$. Moreover, suppose that $P_0\{\|h(\mathbf{X})\|^2 + \|\mathbf{X}\|^2\} < \infty$ and that $P_0\{\pi(\mathbf{X})(1 - \pi(\mathbf{X}))(1, \mathbf{X}^\top)(1, \mathbf{X}^\top)^\top\}$ is positive definite, where $\pi(\mathbf{X}) = \pi(\mathbf{X}; \alpha_0, \boldsymbol{\beta}_0)$. Then, as $N \to \infty$, the following results hold: (1) $\sup_{(\mathbf{x},y)} |\widehat{F}_k(\mathbf{x}, y) - F_k(\mathbf{x}, y)| = o_p(1), k = 0, 1$. (2) The stochastic process $\{\sqrt{n}(\widehat{F}_k(\mathbf{x}, y) - F_k(\mathbf{x}, y)) : (\mathbf{x}, y) \in \mathbb{R}^{p+1}\}$ to a mean zero Gaussian process.

Theorem 5 implies that the proposed empirical likelihood estimators for both $F_0(\mathbf{x}, y)$ and $F_1(\mathbf{x}, y)$ are uniformly consistent and asymptotically normal with a root-n convergence rate, provided that the propensity score model (6) is correctly specified. To construct conformal prediction sets, we impose a parametric working model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ for the common CDF $\Pr(Y(1) \le y \mid \mathbf{X} = \mathbf{x}, D = 0) = \Pr(Y(1) \le y \mid \mathbf{X} = \mathbf{x}, D = 1) = \Pr(Y(1) \le y \mid \mathbf{X} = \mathbf{x})$. Our resampling-based approach for conformal prediction of the potential outcome $Y(1)$ in the control group is described in Algorithm 3. We also show in Theorem 6 that the proposed prediction sets have the desirable double robustness property.

Theorem 6. The prediction sets given by Algorithm 3 maintain an asymptotically correct coverage rate when one of the following two sets of conditions is satisfied: (a) The outcome regression model $F(y \mid \mathbf{x}; \boldsymbol{\theta})$ is correctly specified with $\boldsymbol{\theta}_*$ being the unique true value of $\boldsymbol{\theta}$ and Conditions (A) hold with $P_1\{K^2(\mathbf{X}, Y)\} < \infty$ in place of $P_0\{K(\mathbf{X}, Y)\} < \infty$. (b) The propensity score model (6) is correctly specified, and the matrix $\mathbb{E}\{\pi(\mathbf{X}; \alpha_{**}, \boldsymbol{\beta}_{**})(1 - \pi(\mathbf{X}; \alpha_{**}, \boldsymbol{\beta}_{**}))(1, \mathbf{X}^\top)^\top(1, \mathbf{X}^\top)\}$ is positive definite with $(\alpha_{**}, \boldsymbol{\beta}_{**})$ being the unique true value of $(\alpha, \boldsymbol{\beta})$.

As expected, incorporating additional covariate information from another treatment arm improves estimation of $F_1(\mathbf{x}, y)$. However, our simulation studies show that Algorithm 3 does not exhibit significant improvement in finite-sample setting compared to Algorithm 1 in terms of the length of prediction intervals. This is similar to the fact that, in the context of estimating a mean, a large increase in sample size can significantly reduce the length of the confidence interval for the mean, but a larger sample size usually does not result in a substantial reduction in the length of a prediction interval. The presence of subject-to-subject variation for a future observation means that the reduction in prediction interval length due to increased sample size is typically not substantial.

Thus far, we have discussed the prediction of the potential outcome $Y(1)$ in the control group based on data from the treatment group, using the relationship $dF_0(\mathbf{x}, y) = \exp(-\mathbf{x}^\top \boldsymbol{\beta}) dF_1(\mathbf{x}, y) / \iint \exp(-\mathbf{s}^\top \boldsymbol{\beta}) dF_1(\mathbf{s}, t)$. Prediction of $Y(0)$ in the treatment group can be obtained in a similar way. Additionally, when predicting $Y(1)$ in the general population, we can rely on the relationship $dF(\mathbf{x}, y) = \pi(\mathbf{x})^{-1} dF_1(\mathbf{x}, y) / \iint \pi(\mathbf{s})^{-1} dF_1(\mathbf{s}, t)$. Hence, in Algorithm 3 we bootstrap from $\widehat{F}(\mathbf{x}, y)$ and $\widehat{F}_1(\mathbf{x}, y)$, where $\widehat{F}(\mathbf{x}, y) = \sum_{i=1}^n \widehat{p}_i \widehat{\pi}(\mathbf{X}_i)^{-1} I(\mathbf{X}_i \le \mathbf{x}, Y_i(1) \le y) / \sum_{j=1}^n \widehat{p}_j \widehat{\pi}(\mathbf{X}_j)^{-1}$.

4. Numerical Experiments

In this section, we compare the performance of the proposed and existing conformal inference methods using both simulated and semi-simulated data.

4.1. Performance Evaluation Using Simulated Data

For comparison, we evaluate the performance of the following approaches: (a) MARG, the marginal quantiles of Y; (b) LC, the weighted split-CQR algorithm described in Lei and Candès (2021); (c) COND1, an unsplit conditional approach described

Table 1. Simulated average coverage rate (AC), average length (AL) of prediction intervals and its standard deviation (SD) for an exchangeable outcome at the 95% target level when data are generated from Scenario 1 with prediction estimand $Y(1) \mid D = 0$.

N	Method	AC(%)	AL(SD)	AC(%)	AL(SD)
		(I) both correctly specified		(II) OR correctly specified	
1000	MARG	90.99	71.25 (19.24)	90.99	71.25 (19.24)
	LC	93.81	99.82 (121.65)	93.81	99.82 (121.65)
	COND1	94.72	12.42 (2.66)	94.36	11.80 (2.46)
	COND2	94.64	11.63 (2.37)	94.33	11.57 (2.36)
	BOOT1	94.03	11.94 (2.41)	94.10	11.97 (2.39)
	BOOT2	94.02	11.92 (2.40)	93.88	11.85 (2.39)
2000	MARG	91.30	71.75 (13.90)	91.30	71.75 (13.90)
	LC	92.07	53.59 (14.31)	92.07	53.59 (14.31)
	COND1	94.94	12.22 (2.27)	94.62	11.78 (2.08)
	COND2	94.91	11.77 (2.12)	94.61	11.68 (2.06)
	BOOT1	94.44	11.90 (2.10)	94.42	11.86 (2.03)
	BOOT2	94.44	11.89 (2.08)	94.22	11.76 (2.04)
		(III) PS correctly specified		(IV) both misspecified	
1000	MARG	90.99	71.25 (19.24)	90.99	71.25 (19.24)
	LC	93.81	99.82 (121.65)	93.81	99.82 (121.65)
	COND1	93.91	32.81 (10.73)	93.91	22.62 (4.10)
	COND2	95.49	20.01 (4.04)	95.14	19.35 (3.59)
	BOOT1	93.74	20.69 (3.98)	93.90	19.95 (3.52)
	BOOT2	93.76	20.80 (3.99)	93.71	19.91 (3.58)
2000	MARG	91.30	71.75 (13.90)	91.30	71.75 (13.90)
	LC	92.07	53.59 (14.31)	92.07	53.59 (14.31)
	COND1	94.46	27.08 (6.83)	94.44	20.50 (2.93)
	COND2	95.60	19.99 (3.22)	95.19	19.32 (2.83)
	BOOT1	94.21	20.30 (3.21)	94.29	19.59 (2.78)
	BOOT2	94.23	20.35 (3.13)	94.01	19.42 (2.80)

in Section 2.2 with $S(\mathbf{X}_{n+1}, Y_{n+1}) = 1$; (d) COND2, similar to COND1 but with $S(\mathbf{X}_{n+1}, Y_{n+1}) = 0$; (e) BOOT1, a resampling method based on exponential tilted empirical distribution described in Section 2.3; and (f) BOOT2, a resampling approach incorporating auxiliary information described in Section 3.

For the propensity score (PS) model in our proposed methods, we use logistic regression models as the working models in the first scenario, and we explore the use of machine learning methods in the second scenario. For the outcome regression (OR), we consider the correctly specified Gamma model with shape parameter θ_2 and scale parameter $\theta_2^{-1}\exp(\boldsymbol{\theta}_1^\top\widetilde{\mathbf{X}})$ and a misspecified log-normal model with mean $\boldsymbol{\theta}_1^\top\widetilde{\mathbf{X}}$ and variance θ_2, where $\widetilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$. For the BOOT2 method, we set $h(\mathbf{X}) = \widehat{w}(\mathbf{X})\widehat{\boldsymbol{\theta}}_1^\top\widetilde{\mathbf{X}}$, where $\widehat{w}$ is estimated using logistic regressions or machine learning methods and $\widehat{\boldsymbol{\theta}}_1$ is the MLE under the imposed OR model. As for the LC method, we use quantile random forest (qRF) (Athey, Tibshirani, and Wager 2019) to estimate conditional quantiles and gradient boosting machine (Friedman 2001) to estimate propensity scores, as recommended in Lei and Candès (2021). In each simulation, we generate $N = 1000$ and 2000 data points to construct prediction intervals, evaluate their performance on $M = 5000$ extra data points, and use $B = 2000$ bootstrap samples for resampling-based methods. We repeat the process 5000 times and report summary statistics of average coverage rate (AC) and average length (AL) of the prediction intervals at the 95% level for performance evaluation.

We consider two scenarios: one with right-skewed data (Scenario 1) and the other with approximately symmetric additive errors (Scenario 2). In Scenario 1, the covariate vector $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)^\top$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ with a pairwise correlation coefficient of 0.3. The potential outcome $Y(1)$ is generated from a Gamma distribution with shape parameter 2 and scale parameter $\exp(\boldsymbol{b}^\top\widetilde{\mathbf{X}})/2$ with $\boldsymbol{b} = (0, -1, 1, 1, -1, -1)^\top$. The treatment indicator D is generated from a Bernoulli distribution with a success probability $\{1 + \exp(-\boldsymbol{c}^\top\widetilde{\mathbf{X}} - \gamma X_1 X_2)\}^{-1}$, where $\boldsymbol{c} = (-1, 0.5, 1, 0.5, -0.5, -1)^\top$ and $\gamma = 1$. Four cases of working models are considered: (I) both OR and PS models are correctly specified; (II) only OR model is correct; (III) only PS model is correct; and (IV) both OR and PS models are misspecified.

Table 1 compares the performance of various conformal inference methods in predicting $Y(1)$ in the control arm. Note that results of MARG and LC are the same across different cases as they do not rely on working models. In all cases, the proposed methods, COND1, COND2, BOOT1, and BOOT2, outperform MARG and LC in achieving coverage rates (AC) close to the target rate. Furthermore, the proposed methods exhibit considerably shorter average interval lengths (AL). Among these, COND2 emerges as the most effective approach, striking the balance between adhering to the target coverage rate and maintaining a short average length of the prediction interval. Increasing the sample size does not lead to significant improvements in AC and AL but reduces the variability of the prediction interval length. Interestingly, as the sample size increases, coverage rates for LC decline within each case. This observation corresponds with a notable reduction in interval length for larger sample sizes, a phenomenon not observed with our proposed methods. Furthermore, when both working PS and OR models are accurate (Case I), interval lengths are considerably shorter than in Case (IV), where both models are misspecified. To evaluate the robustness of our proposed methods against PS model misspecification, we compare Case (I) and Case (II), observing similar AC and AL. In contrast, when comparing Case (I) with Case (III), AL nearly doubles when the OR working model is misspecified. Similar observations arise when comparing Case (IV) to Case (II) and Case (III). In summary, for this data generation model, the misspecification of the OR working model has a more substantial impact on interval length than the PS working model. The performance of predicting $Y(1)$ in the entire study population, summarized in Table S1 of the supplementary materials, exhibits patterns similar to those in Table 1.

Our proposed method has a notable advantage over the LC method in its ability to handle various prediction tasks. For illustration, we consider predicting the maximum of future responses, say, $\max_{n+1\le j\le n+5} Y_j(1)$, and compare the performance of our prediction intervals BOOT1 and BOOT2 with the MARG method. The simulation results in Table S3 of the supplementary materials, where a misspecified working OR model and a correctly specified working PS model are used, show that our resampling-based methods significantly outperform the marginal approach by producing shorter prediction intervals while maintaining a coverage rate very close to the target level.

In Scenario 2, we consider approximately symmetric outcome distributions with a high-dimensional covariate $\mathbf{X} = (X_1, \dots, X_{100})^\top$ and use data-adaptive weight $w(\mathbf{x})$ in the construction of prediction intervals. Following Lei and Candès (2021), we generate $Y(1) = 4\left[1+\exp\{-12(X_1-0.5)\}\right]^{-1}\left[1+\exp\{-12(X_2-0.5)\}\right]^{-1} + \epsilon$, where each component of $\mathbf{X}$ is generated from the uniform distribution $U(0,1)$ with $\mathbb{C}\mathrm{orr}(X_j, X_{j'}) = 0.5$, for $j \neq j'$. The error term ϵ follows one of four distributions: $N(0,1)$, $N(0, -\log X_1)$, the t-distribution with 5 degrees of freedom (t_5), and the skew-normal distribution with the location, the scale and the shape parameters being -1, 2, and 3, respectively ($SN(-1,2,3)$). The treatment indicator D is simulated from a Bernoulli distribution with success probability $\{1+B_{24}(X_1)\}/4$, where $B_{24}(\cdot)$ is the CDF of the Beta distribution with parameters $(2,4)$. Here we only compare MARG, LC, COND2, and BOOT1 methods, as the performance of COND1 and BOOT2 are generally inferior to that of COND2 and BOOT1. For the proposed methods, a working OR model with Gaussian error is always assumed and the nonconformity score is based on the Gaussian CDF. For all the methods under comparison, either penalty methods such as LASSO (Tibshirani 1996) and SCAD (Fan and Li 2001) are applied to the OR model with misspecified linear covariate effects, or machine learning methods such as quantile random forest (qRF) and regression forest (RF) (Athey, Tibshirani, and Wager 2019) are, respectively, adopted to estimate the conditional quantiles and conditional means. All the PSs are estimated using the gradient boosting machine (GBM) (Friedman 2001) in this scenario.

Table 2 demonstrates that all methods, except for the proposed conditional approach using RF for the conditional mean in the OR model, perform reasonably well in terms of prediction accuracy and achieve satisfactory average coverage rates when the error distribution does not depart significantly from a symmetric distribution. Our proposed conditional approach COND2 generally has shorter average lengths and is slightly undercovered compared to the LC method. Similar observations can be made from the additional simulation results with low-dimensional covariates presented in the supplementary materials. It is worth pointing out that BOOT1 (RF+GBM) has the best performance among all the methods considered under Scenario 2. This result highlights the merit of the proposed bootstrap approach, which mimics the underlying data generation process by resampling, rather than data splitting, where the sampling weight is estimated using the GBM with statistical guarantees.

4.2. Performance Evaluation Using Semi-Simulated Data

In this section, we evaluate the performance of different conformal inference methods on semi-simulated data from three real-life studies. The first example involves outcome data with approximately symmetric distribution while the other two are related to right-skewed positive data that naturally occur in physical and social science. We start by considering the National Study of Learning Mindsets (NSLM), which is a large randomized controlled trial of an online growth mindset program for school children (Yeager et al. 2019). Our main objective is to predict the potential outcome on treatment $Y(1)$ in the general population. To simulate an observational study, we added heteroscedastic normal errors to nonlinear covariate effects to generate the potential outcome $Y(1)$. More information about the data generation process can be found in section 4.4 of Lei and Candès (2021), which we omit here. We focus on the proposed conditional approach COND2 and the plain bootstrap approach BOOT1, both of which have demonstrated better performance in previous simulation studies. We explore different combinations of working models, including a parametric linear model (LM) $Y = \boldsymbol{\theta}_1^\top \mathbf{X} + \varepsilon$ with $\varepsilon \sim N(0, \theta_2)$ and the qRF for the outcomes, and the logistic regression (LR) and gradient boosting machine (GBM) for propensity scores. Similar to previous findings, the results summarized in Table 3 suggest that all methods perform comparably when the distribution of the error terms is roughly symmetric and without heavy-tailed behavior.

The next example examines the performance of conformal inference methods in the presence of covariate shift using the airfoil dataset available from the UCI Machine Learning Repository (Dua, Dheeru, and Graff, Casey 2019). The dataset includes 1503 measurements of scaled sound pressure level, along with five covariates: (log) frequency, angle of attack, chord length, free-stream velocity, and suction side (log) displacement thickness. It is worth pointing out that the outcome variable in the published dataset is in the log scale after applying the spectral scaling laws (Brooks, Pope, and Marcolini 1989). We randomly split the data in half, with one portion used for training and the other for testing. To simulate covariate shift, we follow Tibshirani et al. (2019) by artificially sampling 25% of the testing data with selection probability inversely proportional to the frequency. More details about the data generation process are available in section 2.3 of their paper. We compare a diverse set of prediction techniques, including modern machine learning approaches and conformal inference methods, that are widely applicable and easy to implement in practice. Specifically, we add OLSp, the classic least-squares prediction interval, and TBCR, the weighted split conformal prediction method proposed by Tibshirani et al. (2019), to the list of methods, in addition to MARG, LC, COND2, and BOOT1. We consider various combinations of outcome regression and covariate-shift models for the proposed and existing conformal prediction methods. We use LM and qRF as working outcome regression models. For

Table 2. Simulated average coverage rate (AC), average length (AL) of prediction intervals and its standard deviation (SD) for an exchangeable outcome at the 95% target level when data are generated from Scenario 2 with prediction estimand $Y(1) \mid D = 0$.

N	Method	AC(%)	AL(SD)	AC(%)	AL(SD)
		$N(0,1)$		$N(0,-\log X_1)$	
1000	MARG	94.47	6.56 (0.22)	93.38	6.03 (0.23)
	LC (LASSO+GBM)	95.42	5.04 (0.53)	94.94	5.39 (0.99)
	LC (SCAD+GBM)	95.39	5.04 (0.53)	94.82	5.43 (1.00)
	LC (qRF+GBM)	95.51	6.14 (0.52)	95.20	5.98 (0.63)
	COND2 (LASSO+GBM)	94.41	4.76 (0.25)	93.92	4.97 (0.42)
	COND2 (SCAD+GBM)	94.72	4.85 (0.26)	94.20	5.12 (0.43)
	COND2 (RF+GBM)	92.97	4.32 (0.22)	92.63	4.37 (0.35)
	BOOT1 (LASSO+GBM)	97.16	5.44 (0.25)	95.67	5.54 (0.37)
	BOOT1 (SCAD+GBM)	96.69	5.32 (0.26)	95.26	5.47 (0.37)
	BOOT1 (RF+GBM)	95.72	4.85 (0.20)	94.64	4.78 (0.30)
2000	MARG	94.77	6.60 (0.16)	93.60	6.05 (0.16)
	LC (LASSO+GBM)	95.20	4.89 (0.36)	94.71	5.19 (0.64)
	LC (SCAD+GBM)	95.13	4.90 (0.36)	94.81	5.29 (0.66)
	LC (qRF+GBM)	95.31	5.90 (0.35)	95.01	5.71 (0.35)
	COND2 (LASSO+GBM)	94.66	4.75 (0.17)	94.17	4.99 (0.29)
	COND2 (SCAD+GBM)	94.78	4.80 (0.17)	94.31	5.10 (0.29)
	COND2 (RF+GBM)	92.94	4.04 (0.14)	92.81	4.21 (0.24)
	BOOT1 (LASSO+GBM)	95.98	5.05 (0.16)	94.87	5.25 (0.26)
	BOOT1 (SCAD+GBM)	95.77	5.03 (0.16)	94.71	5.24 (0.26)
	BOOT1 (RF+GBM)	95.83	4.57 (0.13)	94.71	4.66 (0.23)
		t_5		$SN(-1,2,3)$	
1000	MARG	94.51	7.27 (0.33)	94.41	7.35 (0.30)
	LC (LASSO+GBM)	95.38	6.24 (0.94)	95.41	6.18 (0.74)
	LC (SCAD+GBM)	95.44	6.25 (0.98)	95.26	6.16 (0.81)
	LC (qRF+GBM)	95.39	7.11 (0.92)	95.52	6.99 (0.54)
	COND2 (LASSO+GBM)	94.64	5.82 (0.43)	94.45	5.78 (0.33)
	COND2 (SCAD+GBM)	94.91	5.92 (0.43)	94.82	5.87 (0.32)
	COND2 (RF+GBM)	93.66	5.44 (0.39)	92.81	5.34 (0.28)
	BOOT1 (LASSO+GBM)	96.61	6.57 (0.39)	97.19	6.64 (0.33)
	BOOT1 (SCAD+GBM)	96.36	6.46 (0.42)	96.80	6.49 (0.34)
	BOOT1 (RF+GBM)	95.39	5.98 (0.35)	95.81	6.02 (0.27)
2000	MARG	94.77	7.31 (0.24)	94.65	7.39 (0.21)
	LC (LASSO+GBM)	95.17	5.96 (0.58)	95.17	5.95 (0.48)
	LC (SCAD+GBM)	95.18	5.98 (0.58)	95.13	5.97 (0.51)
	LC (qRF+GBM)	95.22	6.77 (0.57)	95.25	6.75 (0.38)
	COND2 (LASSO+GBM)	94.80	5.78 (0.28)	94.67	5.78 (0.22)
	COND2 (SCAD+GBM)	94.92	5.84 (0.28)	94.85	5.84 (0.22)
	COND2 (RF+GBM)	93.55	5.13 (0.26)	92.68	5.07 (0.19)
	BOOT1 (LASSO+GBM)	95.69	6.11 (0.26)	96.00	6.16 (0.21)
	BOOT1 (SCAD+GBM)	95.59	6.09 (0.26)	95.89	6.13 (0.21)
	BOOT1 (RF+GBM)	95.42	5.72 (0.25)	95.83	5.75 (0.18)

Table 3. Simulated average coverage rate (AC), average length (AL) of prediction intervals and its standard deviation (SD) for the NSLM data.

Method	AC(%)	AL(SD)
MARG	94.20	2.10 (0.13)
LC (LM+LR)	95.65	2.06 (0.28)
LC (qRF+GBM)	95.55	2.24 (0.28)
COND2 (LM+LR)	94.38	1.89 (0.13)
COND2 (LM+GBM)	94.40	1.89 (0.13)
BOOT1 (LM+LR)	94.70	1.91 (0.11)
BOOT1 (LM+GBM)	94.68	1.91 (0.11)

the working covariate-shift models, we use the LR and GBM as weighted models and consider an unweighted (UW) model as well, following the simulation setting in Tibshirani et al. (2019). The left panel of Table 4 summarizes the performance of various conformal inference procedures. It is worth noting that the average interval length (AL) summary measure is not always available for the TBCR algorithm as their algorithm may fail to produce valid quantiles. Thus, in addition to AL, we also report the median interval length (ML) to provide a more comprehensive evaluation. As expected, OLSp and its split counterpart, TBCR (LM+UW), have poor performance and suffer from severe under-coverage due to their failure to account for covariate shift. In contrast, all other methods exhibit comparable performance. Specifically, COND2 achieves the target coverage rates for both combinations (i.e., LM+LR and LM+GBM), while BOOT1 results in the shortest prediction interval length. However, it is worth noting that BOOT1 falls slightly short of achieving the exact 95% target level by about 0.2%.

Finally, we consider the well-known Boston Housing Prices dataset (Harrison and Rubinfeld 1978), which is available from the R package MASS and provides information on housing and demographic factors in Boston suburbs. In this dataset, the outcome variable is the median value of owner-occupied homes in a Boston suburb or town. We randomly split the data into training and testing sets, with 75% used for training and 25% for testing, so that there is no covariate shift between the two sets. Following equation (A.1) in Harrison and Rubinfeld (1978), we adopt a linear model for the log-transformed outcomes, in addition to the qRF, as working OR models. We use LR and

Table 4. Simulated average coverage rate (AC), average length (AL), median length (ML) of prediction intervals and their standard deviation (SD) for the airfoil data and the Boston housing prices data (price in thousands).

Method	AC(%)	ML(SD)	AL(SD)	AC(%)	ML(SD)	AL(SD)
	airfoil data			Boston housing prices data		
MARG	94.59	26.03 (0.49)	26.03 (0.49)	96.30	41.57 (0.62)	41.57 (0.62)
OLSp	89.51	19.85 (0.44)	19.86 (0.44)	94.29	15.27 (0.76)	16.58 (0.78)
TBCR (LM+UW)	89.35	20.01 (1.37)	20.01 (1.37)	95.37	18.47 (2.97)	19.96 (3.17)
TBCR (LM+LR)	95.67	27.28 (4.20)	–	95.53	18.85 (4.22)	–
TBCR (LM+GBM)	95.95	27.80 (4.66)	–	97.14	23.34 (5.37)	–
LC (LM+LR)	95.93	29.23 (6.57)	30.14 (5.18)	95.84	17.78 (3.47)	19.02 (3.47)
LC (LM+GBM)	96.20	29.66 (6.71)	30.52 (5.15)	97.72	26.58 (10.25)	27.36 (8.49)
LC (qRF+GBM)	96.42	25.16 (2.88)	24.90 (2.57)	98.36	29.63 (3.40)	29.61 (3.10)
COND2 (LM+LR)	95.34	26.41 (2.33)	26.86 (2.01)	94.77	16.44 (1.55)	17.81 (1.67)
COND2 (LM+GBM)	95.37	26.58 (2.42)	27.07 (1.99)	95.39	17.45 (2.02)	19.12 (2.12)
BOOT1 (LM+LR)	94.82	24.73 (1.66)	24.75 (1.64)	94.84	16.66 (1.42)	18.05 (1.52)
BOOT1 (LM+GBM)	94.82	24.78 (1.79)	24.80 (1.77)	94.60	16.56 (2.04)	17.94 (2.19)

GBM for the working covariate-shift models. Noting that no transformation is applied to the outcomes when applying the qRF as a working OR model in this example. All the prediction methods mentioned in the previous example (i.e., MARG, OLSp, TBCR, LC, COND2, and BOOT1) are considered here for comparison. The right panel of Table 4 shows that the OLSp method performs comparably to the unweighted TBCR method. Data splitting helps achieve the exact target coverage rate, but it also leads to longer interval lengths and worsened prediction stability, as indicated by the standard deviation. Simple parametric working models demonstrate better performance in this simulation study. In contrast, methods based on qRF and/or GBM tend to produce longer prediction intervals with coverage rates much higher than the target 95% level. The proposed unsplit conditional approach and resampling method also demonstrate satisfactory performance. Among all the methods considered, including the split methods, LC (LM+LR) and COND2 (LM+GBM) are the top-performing methods, achieving the target coverage rate while maintaining the shortest average interval length.

Supplementary Materials

The online supplement contains proofs of Theorems 1–6, the proposed conformal prediction algorithms for the maximum of several responses in the presence of covariate shift, details of a proposed jackknife and an alternative resampling-based conformal prediction method in the absence of covariate shift, and additional simulation results.

Acknowledgments

The authors would like to thank the editor, associate editor, and two reviewers for their insightful comments which have helped improve the manuscript substantially.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Liu's research is supported by the National Key R&D Program of China (2021YFA1000100 and 2021YFA1000101), the National Natural Science Foundation of China (12171157, 71931004, 32030063), Fundamental Research Funds for the Central Universities and the 111 Project (B14019).

References

Angelopoulos, A. N., and Bates, S. (2022), "Conformal Prediction: A Gentle Introduction," *Foundations and Trends in Machine Learning*, 16, 494–591. [1]

Athey, S., Tibshirani, J., and Wager, S. (2019), "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178. [9,10]

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a), "The Limits of Distribution-Free Conditional Predictive Inference," *Information and Inference: A Journal of the IMA*, 10, 455–482. [1]

——— (2021b), "Predictive Inference with the Jackknife+," *The Annals of Statistics*, 49, 486–507. [1,2]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1]

Brooks, T. F., Pope, D. S., and Marcolini, M. A. (1989), "Airfoil Self-Noise and Prediction," Technical Report 1218, NASA. [10]

Candès, E., Lei, L., and Ren, Z. (2023), "Conformalized Survival Analysis," *Journal of the Royal Statistical Society*, Series B, 85, 24–45. [1,3]

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021), "Distributional Conformal Prediction," *Proceedings of the National Academy of Sciences*, 118, e2107794118. [1]

Cox, D. R. (1975), "Prediction Intervals and Empirical Bayes Confidence Intervals," *Journal of Applied Probability*, 12, 47–55. [4]

Dua, Dheeru and Graff, Casey (2019), "UCI Machine Learning Repository," available at *http://archive.ics.uci.edu/ml*. [10]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [10]

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. (2022), "Conformal Prediction Under Feedback Covariate Shift for Biomolecular Design," *Proceedings of the National Academy of Sciences*, 119, e2204569119. [1]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [1,9,10]

Harris, I. R. (1989), "Predictive Fit for Natural Exponential Families," *Biometrika*, 76, 675–684. [4]

Harrison, D., and Rubinfeld, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [11]

Jin, Y., and Candès, E. J. (2023), "Selection by Prediction with Conformal P-values," *Journal of Machine Learning Research*, 24, 1–41. [1]

Kim, B., Xu, C., and Barber, R. (2020), "Predictive Inference is Free with the Jackknife+-After-Bootstrap," in *Advances in Neural Information Processing Systems* (Vol. 33), pp. 4138–4149. [2]

Lawless, J. F., and Fredette, M. (2005), "Frequentist Prediction Intervals and Predictive Distributions," *Biometrika*, 92, 529–542. [4]

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. [1,2]

Lei, L., and Candès, E. J. (2021), "Conformal Inference of Counterfactuals and Individual Treatment Effects," *Journal of the Royal Statistical Society*, Series B, 83, 911–938. [1,3,8,9,10]

Owen, A. B. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120. [7]

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002), "Inductive Confidence Machines for Regression," in *European Conference on Machine Learning*, Springer, pp. 345–356. [2]

Park, S., Dobriban, E., Lee, I., and Bastani, O. (2022), "PAC Prediction Sets Under Covariate Shift," in *International Conference on Learning Representations*. [1]

Qin, J. (2017), *Biased Sampling, Over-Identified Parameter Problems and Beyond*, Singapore: Springer. [4,7]

Qin, J., Liu, Y., and Li, P. (2023), "A Selective Review of Statistical Methods Using Calibration Information from Similar Studies," *Statistical Theory and Related Fields*, 6, 175–190. [7]

Qiu, H., Dobriban, E., and Tchetgen Tchetgen, E. (2023), "Prediction Sets Adaptive to Unknown Covariate Shift," *Journal of the Royal Statistical Society*, Series B, 85, 1680–1705. [1]

Romano, Y., Patterson, E., and Candès, E. J. (2019), "Conformalized Quantile Regression," in *Advances in Neural Information Processing Systems* (Vol. 32), pp. 3543–3553. [1]

Shafer, G., and Vovk, V. (2008), "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research*, 9, 371–421. [1,2]

Tian, Q., Meng, F., Nordman, D. J., and Meeker, W. Q. (2022), "Predicting the Number of Future Events," *Journal of the American Statistical Association*, 117, 1296–1310. [4]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [10]

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019), "Conformal Prediction Under Covariate Shift," in *Advances in Neural Information Processing Systems* (Vol. 32), pp. 2530–2540. [1,2,10,11]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [3]

Vovk, V. (2015), "Cross-Conformal Predictors," *Annals of Mathematics and Artificial Intelligence*, 74, 9–28. [2]

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

Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024), "Doubly Robust Calibration of Prediction Sets Under Covariate Shift," *Journal of the Royal Statistical Society, Series B*, 1–23. [1]

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., et al. (2019), "A National Experiment Reveals Where a Growth Mindset Improves Achievement," *Nature*, 573, 364–369. [10]

Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2024), "Conformal Sensitivity Analysis for Individual treatment Effects," *Journal of the American Statistical Association*, 119, 122–135. [1]