Conformalized Survival Analysis

Emmanuel J. Candès[1,2], Lihua Lei[1], and Zhimei Ren[1]

[1]Department of Statistics, Stanford University, Stanford, CA 94305
[2]Department of Mathematics, Stanford University, Stanford, CA 94305

March 18, 2021

Abstract

Existing survival analysis techniques heavily rely on strong modelling assumptions and are, therefore, prone to model misspecification errors. In this paper, we develop an inferential method based on ideas from conformal prediction, which can wrap around any survival prediction algorithm to produce calibrated, covariate-dependent lower predictive bounds on survival times. In the Type I right-censoring setting, when the censoring times are completely exogenous, the lower predictive bounds have guaranteed coverage in finite samples without any assumptions other than that of operating on independent and identically distributed data points. Under a more general conditionally independent censoring assumption, the bounds satisfy a doubly robust property which states the following: marginal coverage is approximately guaranteed if either the censoring mechanism or the conditional survival function is estimated well. Further, we demonstrate that the lower predictive bounds remain valid and informative for other types of censoring. The validity and efficiency of our procedure are demonstrated on synthetic data and real COVID-19 data from the UK Biobank.

Keywords. Censoring, survival time, prediction interval, weighted conformal inference, distribution boosting, random forests.

1 Introduction

The COVID-19 pandemic has placed extraordinary demands on health systems (e.g., Ranney et al., 2020). In turn, these demands create an unavoidable need for medical resource allocation and, in response, several groups of researchers have communicated clinical ethics recommendations (e.g., Emanuel et al., 2020; Vergano et al., 2020). By and large, these recommendations require a reliable benefit assessment of receiving specific types of medical resources; see Table 2 of Emanuel et al. (2020). Clearly, one benefit measure of interest might be the survival time, the time lapse between the confirmation of COVID-19 and an event such as death or reaching a critical state, should this ever occur.

1.1 Survival analysis

Survival times are not always observed due to censoring (Leung et al., 1997). A main goal of survival analysis is to infer the survival function—the probability that a patient will survive beyond any specified time—from censored data. The Kaplan-Meier curve (Kaplan and Meier, 1958) produces such an inference when the population under study is a group of patients with certain characteristics. On the positive side, the Kaplan-Meier curve does not make any assumption on the distribution of survival times. On the negative side, it can only be applied to a handful of subpopulations because it requires sufficiently many events in each subgroup (Kalbfleisch and Prentice, 2011). More often than not, the scientist has available multiple categorical and continuous covariates, and it thus becomes of interest to understand heterogeneity by studying the conditional survival function; that is, the dependence on the available factors. In the conditional setting, however, distribution-free inference for the conditional survival function gets to be challenging. Standard approaches make parametric or nonparametric assumptions about the distribution of the covariates and that of the survival times conditional on covariate values. A well-known example is of course the celebrated Cox model which posits a proportional hazards model in which an unspecified nonparametric base line is modified via a parametric model describing how the hazard varies in response to explanatory covariates (Cox, 1972; Breslow, 1975). Other popular models, such as accelerated failure time (AFT) (Cox, 1972; Wei, 1992) and proportional odds models (Murphy et al., 1997; Harrell Jr, 2015), also combine nonparametric and parametric model specifications.

As medical technologies produce ever larger and more complex clinical datasets, we have witnessed a rapid development of machine learning methods adapted to high-dimensional and heterogeneous survival data (e.g., Verweij and Van Houwelingen, 1993; Faraggi and Simon, 1995; Tibshirani, 1997; Gui and Li, 2005; Hothorn et al., 2006; Zhang and Lu, 2007; Ishwaran et al., 2008; Witten and Tibshirani, 2010; Goeman, 2010; Simon et al., 2011; Katzman et al., 2016; Lao et al., 2017; Wang et al., 2019; Li and Bradic, 2020). An appealing feature of these methods is that they typically do not make modeling assumptions.[1] The downside is that it is often challenging to quantify the uncertainty for these methods. To be sure, blind application of off-the-shelf uncertainty quantification tools, such as the bootstrap (Efron, 1979; Efron and Tibshirani, 1994), can yield unreliable results since their validity 1) rests on implicit modeling assumptions, and 2) holds only asymptotically (e.g., Lei and Candès, 2020; Ratkovic and Tingley, 2021).

1.2 Prediction intervals

For decision-making in sensitive and uncertain environments—think of the COVID-19 pandemic—it is preferable to produce prediction intervals for the *uncensored* survival time with guaranteed coverage rather than point predictions. In this regard, the use of $(1-\alpha)$ prediction intervals is an effective way of summarizing what can be learned from the available data; wide intervals reveal a lack of knowledge and keep overconfidence at arm's length. Here and below, an interval is said to be a $(1-\alpha)$ prediction interval if it has the property that it contains the true label, here, the survival time, at least $100(1-\alpha)\%$ of the time (a formal definition is in Section 2). Prediction intervals have been widely studied in statistics (e.g., Wilks, 1941; Wald, 1943; Aitchison and Dunsmore, 1980; Stine, 1985; Geisser, 1993; Vovk et al., 2005; Krishnamoorthy and Mathew, 2009) and much research has been concerned with the construction of covariate-dependent intervals.

Of special interest is the subject of conformal inference, a generic procedure that can be used in conjunction with sophisticated machine learning prediction algorithms to produce prediction intervals with valid marginal coverage without making any distributional assumption whatsoever (e.g., Saunders et al., 1999; Vovk, 2002; Vovk et al., 2005; Lei and Wasserman, 2014; Tibshirani et al., 2019). While coverage is only guaranteed in a marginal sense, it has been empirically observed that some conformal prediction methods can also achieve near conditional coverage—that is, coverage assuming a fixed value of the covariates—when some key parameters of the underlying conditional distribution can be estimated reasonably well (e.g., Sesia and Candès, 2020; Lei and Candès, 2020).

1.3 Our contribution

Standard conformal inference requires fully observed outcomes and is not directly applicable to samples with censored outcomes. In this paper, we extend conformal inference to handle right-censored

[1] To quote from Efron (2020): " Neither surface nor noise is required as input to randomForest, gbm, or their kin.

outcomes in the setting of Type-I censoring (e.g., Leung et al., 1997). This setting assumes that the censoring time is observed for every unit while the outcome is only observed for uncensored units. In particular, we generate a covariate-dependent lower predictive bound (LPB) on the uncensored survival time, which can be regarded as a one-sided $(1-\alpha)$-prediction interval. As we just argued, the LPB is a conservative assessment of the survival time, which is particularly desirable for high-stakes decision-making. A low LPB value suggests either a high risk for the patient, or a high degree of uncertainty for similar patients due to data scarcity. Either way, the signal to a decision-maker is that the patient deserves some attention.

Under the completely independent censoring assumption defined below, which states that the censoring time is independent of both the outcome and covariates, our LPB provably yields a $(1-\alpha)$ prediction interval. This property holds in finite samples *without any assumption other than that of operating on i.i.d. samples*. Under the more general conditionally independent censoring assumption introduced later, our LPB satisfies a *doubly robust* property which states the following: marginal coverage is approximately guaranteed if either the censoring mechanism or the conditional survival function is estimated well. In the latter case, the LPB even has approximately guaranteed conditional coverage.

Readers familiar with conformal inference would notice that the above guarantees can be achieved by simply applying conformal inference to the censored outcomes, i.e., by constructing an LPB on the censored outcome treated as the response. This unsophisticated approach is conservative. Instead, we will see how to provide tighter bounds and sharper inference by applying conformal inference on a subpopulation with large censoring times; that is, on which censored outcomes are closer to actual outcomes. To achieve this, we shall see how to carefully combine the selection of a subpopulation with ideas from weighted conformal inference (Tibshirani et al., 2019).

Lastly, while we focus on clinical examples, it will be clear from our exposition that our methods can be applied to other time-to-event outcomes in a variety of other disciplines, such as industrial life testing (Bain, 2017), sociology (Allison, 1984), and economics (Powell, 1986; Hong and Tamer, 2003; Sant'Anna, 2016).

2 Prediction intervals for survival times

2.1 Problem setup

Let X_i, C_i, T_i, $i = 1, \ldots, n$, be respectively the vector of covariates, the censoring time, and the survival time of the i-th unit/patient. Throughout the paper, we assume that (X_i, C_i, T_i) are i.i.d. copies of the random vector (X, C, T). We consider the Type I right-censoring setting, where the observables for the i-th unit include X_i, C_i, and the censored survival time $\widetilde{T}_i$, defined as the minimum of the survival and censoring time:

$$\widetilde{T}_i = \min(T_i, C_i).$$

For instance, if T_i measures the time lapse between the admission into the hospital and death, and C_i measures the time lapse between the admission into the hospital and the day data analysis is conducted, then $\widetilde{T}_i = T_i$ if the i-th patient died before the day of data analysis and $\widetilde{T}_i = C_i$ if she survives beyond that day.

The censoring time C partially masks information from the inferential target T. As discussed by Leung et al. (1997), it is necessary to impose constraints on the dependence structure between T and C to enable meaningful inference. In particular, we make the following **conditionally independent censoring assumption** (e.g., Kalbfleisch and Prentice, 2011):

$$T \perp\!\!\!\perp C \mid X. \tag{1}$$

This assumes away any unmeasured confounder affecting both the survival and censoring time; please see immediately below for an example. In some cases, we also consider the **completely independent censoring assumption**, which is stronger in the sense that it implies the former:

$$(T, X) \perp\!\!\!\perp C. \tag{2}$$

For instance, in a randomized clinical trial, the end-of-study censoring time C is defined as the time lapse between the recruitment and the end of the study. For single-site trials, C is often modelled as a draw from an exogenous stochastic process (e.g., Carter, 2004; Gajewski et al., 2008) and thus obeys (2). For multicentral trials, C is often assumed to depend on the site location only (e.g., Carter et al., 2005; Anisimov and Fedorov, 2007; Barnard et al., 2010), and thus (1) holds as soon as the vector of covariates includes the site of the trial.

Although (1) is a strong assumption, it is a widely used starting point to study survival analysis methods (Kalbfleisch and Prentice, 2011). We leave the investigation of informative censoring (e.g., Lagakos, 1979; Wu and Carroll, 1988; Scharfstein and Robins, 2002) to future research. Additionally, whereas the setting of Type I censoring appears to be restrictive, we will show in Section 2.4 that an LPB in this setting can still be informative for other censoring types.

2.2 Naive lower predictive bounds

Our ultimate goal is to generate a covariate-dependent LPB as a conservative assessment of the uncensored survival time T. Denote by $\hat{L}(\cdot)$ a generic LPB estimated from the observed data $(X_i, C_i, \widetilde{T}_i)_{i=1}^n$. We say an LPB is *calibrated* if it satisfies the following coverage criterion:

$$\mathbb{P}\left(T \geq \hat{L}(X)\right) \geq 1 - \alpha, \tag{3}$$

where α is a pre-specified level (e.g., 0.1), and the probability is computed over both $\hat{L}(\cdot)$ and a future unit (X, C, T) that is independent of $(X_i, C_i, T_i)_{i=1}^n$.

Since $\widetilde{T} \leq T$, any calibrated LPB on the censored survival time $\widetilde{T}$ is also a calibrated LPB on the uncensored survival time T. Consequently, a naive approach is to discard the censoring time C_i's and construct an LPB on $\widetilde{T}$ directly. Since the samples $(X_i, \widetilde{T}_i)$ are i.i.d., a distribution-free calibrated LPB on $\widetilde{T}$ can be obtained via standard techniques from conformal inference (e.g., Vovk et al., 2005; Lei et al., 2018; Romano et al., 2019b). Our first result is somewhat negative: indeed, it states that all distribution-free calibrated LPBs on T must be LPBs on $\widetilde{T}$.

Theorem 1. *Take $X \in \mathbb{R}^p$ and $C \geq 0$, $T \geq 0$. Assume that $\hat{L}(\cdot)$ is a calibrated LPB on T for all joint distributions of (X, C, T) obeying the conditionally independent censoring assumption with X being continuous and (T, C) being continuous or discrete.*[2] *Then for any such distribution,*

$$\mathbb{P}(\widetilde{T} \geq \hat{L}(X)) \geq 1 - \alpha.$$

An LPB constructed by taking $\widetilde{T}$ as the response may be calibrated but also overly conservative because of the censoring mechanism. To see this, note that the oracle LPB on $\widetilde{T}$ is, by definition, the α-th conditional quantile of $\widetilde{T} \mid X$, denoted by $\tilde{q}_\alpha(X)$. Similarly, let $q_\alpha(X)$ be the oracle LPB on T. Under the conditionally independent censoring assumption,

$$\begin{aligned}\mathbb{P}(T \geq q_\alpha(x) \mid X = x) = 1 - \alpha &= \mathbb{P}(\widetilde{T} \geq \tilde{q}_\alpha(x) \mid X = x) \\ &= \mathbb{P}(T \geq \tilde{q}_\alpha(x) \mid X = x)\mathbb{P}(C \geq \tilde{q}_\alpha(x) \mid X = x).\end{aligned}$$

[2] Our proof can be extended to include the case where either C or T or both are mixtures of discrete and continuous distributions but we do not consider such extensions here.

If the censoring times are small, the gap between $\tilde{q}_\alpha(x)$ and $q_\alpha(x)$ can be large. For illustration, assume that X, C, and T are mutually independent, and $T \sim \text{Exp}(1), C \sim \text{Exp}(b)$. It is easy to show that $q_\alpha(X) = -\log(1-\alpha)$ and $\tilde{q}_\alpha(X) = -\log(1-\alpha)/(1+b)$. Thus, a naive approach taking $\widetilde{T}$ as a target of inference can be arbitrarily conservative.

In sum, Theorem 1 implies that any calibrated LPB on T must be a calibrated LPB on $\widetilde{T}$, under the conditionally independent censoring assumption only. This is why to make progress and overcome the limitations of the naive approach, we shall need additional distributional assumptions.

2.3 Leveraging the censoring mechanism

We have just seen that the conservativeness of the naive approach is driven by small censoring times. A heuristic way to mitigate this issue is to discard units with small values of C. Consider a threshold c_0, and extract the subpopulation on which $C \geq c_0$. One immediate issue with this is that the selection induces a distributional shift between the subpopulation and the whole population, namely,

$$(X, C, T) \overset{\text{d}}{\neq} (X, C, T) \mid C \geq c_0.$$

For instance, the patients with larger censoring times tend to be healthier than the remaining ones. To examine the distributional shift in details, note that the joint distribution of $(X, \widetilde{T})$ on the whole population is $P_X \times P_{\widetilde{T}|X}$ while that on the subpopulation is

$$P_{(X,\widetilde{T})|C \geq c_0} = P_{X|C \geq c_0} \times P_{\widetilde{T}|X, C \geq c_0}.$$

Next, observe that $P_{\widetilde{T}|X,C\geq c_0} \neq P_{\widetilde{T}|X}$ even under the completely independent censoring assumption because $(T, X) \perp\!\!\!\perp C$ does not imply $\widetilde{T} \perp\!\!\!\perp C \mid X$ in general. For example, as in Section 2.2, if X, C, and T are mutually independent and $T, C \overset{i.i.d.}{\sim} \text{Exp}(1)$, then $\mathbb{P}(\widetilde{T} \geq a, C \geq a) = \mathbb{P}(\widetilde{T} \geq a) > \mathbb{P}(\widetilde{T} \geq a)\mathbb{P}(C \geq a)$, for any $a > 0$. As a result, both the covariate distribution and the conditional distribution of $\widetilde{T}$ given X differ in the two populations.

Now consider a secondary censored outcome $\widetilde{T} \wedge c_0$, where $a \wedge b = \min\{a, b\}$. We have

$$\begin{aligned} P_{(X,\widetilde{T}\wedge c_0)|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}\wedge c_0|X,C\geq c_0} &\overset{(a)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X,C\geq c_0} \\ &\overset{(b)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X}, \end{aligned} \tag{4}$$

where (a) uses the fact that

$$T \wedge c_0 = \widetilde{T} \wedge c_0, \quad \text{if } C \geq c_0,$$

and (b) follows from the conditionally independent censoring assumption. On the other hand, the joint distribution of $(X, T \wedge c_0)$ on the whole population is

$$P_{(X,T\wedge c_0)} = P_X \times P_{T\wedge c_0|X}. \tag{5}$$

Contrasting (4) with (5), we observe that *there is only a covariate shift* between the subpopulation and the whole population.

The likelihood ratio between the two covariate distributions is

$$\frac{dP_X}{dP_{X|C\geq c_0}}(x) = \frac{\mathbb{P}(C \geq c_0)}{\mathbb{P}(C \geq c_0 \mid X = x)}. \tag{6}$$

Applying the one-sided version of weighted conformal inference (Tibshirani et al., 2019), discussed in the next section, gives a calibrated LPB on $T \wedge c_0$, and thus a calibrated LPB on T. With sufficiently many units with large values of C, we can choose a large threshold c_0 to reduce the loss of power caused by censoring. We emphasize that there is no contradiction with Theorem 1 because, as shown in Section 3, weighted conformal inference requires $\mathbb{P}(C \geq c_0 \mid X)$ to be (approximately) known.

We refer to the denominator $\mathbb{P}(C \geq c_0 \mid X = x)$ in (6) as the *censoring mechanism*, denoted by $c(x; c_0)$. We write it as $c(x)$ for brevity when no confusion can arise. This is the conditional survival function of C evaluated at c_0. Under a censoring of Type I, the C_i's are fully observed while the T_i's are only partially observed. Thus, $\mathbb{P}(C \mid X)$ is typically far easier to estimate than $\mathbb{P}(T \mid X)$. Practically, the censoring mechanism is usually far better understood than the conditional survival function of T; for example, as mentioned in Section 2.1, in randomized clinical trials, C often solely depends on the site location.

Under the completely independent censoring assumption, the covariate shift even disappears since $P_X = P_{X|C \geq c_0}$. In this case, we can apply a one-sided version of conformal inference to obtain a calibrated LPB on $T \wedge c_0$, and hence a calibrated LPB on T (e.g., Vovk et al., 2005; Lei et al., 2018; Romano et al., 2019b). With infinite samples, as $c_0 \to \infty$, the method is tight in the sense that the censoring issue disappears. Again, this result does not contradict Theorem 1, which requires the LPB to be calibrated under the weaker condition (1). With finite samples, there is a tradeoff between the choice of the threshold c_0 and the size of the induced subpopulation.

2.4 Beyond Type-I censoring

In practice, censoring can be driven by multiple factors. As discussed in Leung et al. (1997), the two most common types of right censoring in a clinical study are the end-of-study censoring caused by the trial termination and the loss-to-follow-up censoring caused by unexpected attrition. Let C_{end} denote the former and C_{loss} the latter. By definition, C_{end} is observable for every patient, as long as the entry times are accurately recorded. When the event is not death, C_{loss} is observable for everyone as well. However, when the event is death, C_{loss} can only be observed for surviving patients. This is because for dead patients, it is impossible to know when they would have been lost to follow-up, had they survived.

In survival analysis without loss-to-follow-up censoring, or time-to-event analysis with non-death events, the setting of Type I censoring considered in this paper is plausible. However, it is found that both the end-of-study and loss-to-follow-up censoring are involved in many applications (Leung et al., 1997). In these cases, the effective censoring time C is the minimum of C_{end} and C_{loss}, and is only observable for surviving patients, namely the patients with $T > C$. This situation prevents us from applying Algorithm 1 below because the subpopulation with $C \geq c_0$ is not fully observed. If we use the subpopulation whose C is 1) observed and 2) larger than or equal to a threshold c_0 instead, then the joint distribution of (X, T) becomes $P_{X|C \geq c_0, T > C} \times P_{T|X, C \geq c_0, T > C}$. The extra conditioning event $T > C$ induces a shift of the conditional distribution, since $P_{T|X, C \geq c_0, T > C} \neq P_{T|X, C \geq c_0}$ in general, rendering the weighted split conformal inference invalid.

Our method can nevertheless be adapted to yield meaningful inference under an additional assumption:

$$(T, C_{\text{loss}}) \perp\!\!\!\perp C_{\text{end}} \mid X. \tag{7}$$

The assumption (7) is often plausible since the randomness of the end-of-study censoring time often comes from the entry time of a patient, which is arguably exogenous to the survival time and attrition, at least when conditioning on a few demographic variables. Let $T' = T \wedge C_{\text{loss}}$, the survival time censored merely by the loss to follow-up. Then the censored survival time $\tilde{T} = T \wedge C = T' \wedge C_{\text{end}}$, and (7) implies that $T' \perp\!\!\!\perp C_{\text{end}} \mid X$, an analog of the conditionally independent censoring assumption (1). Since C_{end} is observed for every patient, Algorithm 1 can be applied to produce an LPB $\hat{L}(\cdot)$ such that

$$\mathbb{P}(T' \geq \hat{L}(X)) \geq 1 - \alpha \Longrightarrow \mathbb{P}(T \geq \hat{L}(X)) \geq 1 - \alpha.$$

An observation in conjunction with this line of reasoning is that unlike most survival analysis techniques, our method distinguishes two sources of censoring and takes advantage of the censoring mechanism. It can be regarded as a building block to remove the adverse effect of C_{end}. It remains an interesting question whether the censoring issue induced by C_{loss} can be resolved or alleviated.

3 Conformal inference for censored outcomes

3.1 Weighted conformal inference

Returning to (4) and (5), the goal is to construct an LPB $\hat{L}(\cdot)$ on $T \wedge c_0$ from training samples $(X_i, \tilde{T}_i \wedge c_0)_{C_i \geq c_0} = (X_i, T_i \wedge c_0)_{C_i \geq c_0}$ such that

$$\mathbb{P}(T \wedge c_0 \geq \hat{L}(X)) \geq 1 - \alpha.$$

Since $T \wedge c_0 \leq T$, $\hat{L}(\cdot)$ is a calibrated LPB on T. We consider c_0 to be a fixed threshold in Section 3.1 and 3.2, and discuss a data-adaptive approach to choosing this threshold in Section 3.3.

To deal with covariate shifts, Tibshirani et al. (2019) introduced weighted conformal inference, which extends standard conformal inference (e.g., Vovk et al., 2005; Shafer and Vovk, 2008; Lei and Wasserman, 2014; Barber et al., 2019a,b; Sadinle et al., 2019; Romano et al., 2020; Cauchois et al., 2020)). Imagine we have i.i.d. training samples $(X_i, Y_i)_{i=1}^n$ drawn from a distribution $P_X \times P_{Y|X}$ and wish to construct predictive intervals for test points drawn from the target distribution $Q_X \times P_{Y|X}$ (in standard conformal inference, $P_X = Q_X$). Assuming $w(x) = dQ_X(x)/dP_X(x)$ is known, then weighted conformal inference produces predictive intervals $\hat{C}(\cdot)$ with the property

$$\mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}\left(Y \in \hat{C}(X)\right) \geq 1 - \alpha.$$

Above, the probability is computed over both the training set and the test point (X, Y). In our case, the outcome is $T \wedge c_0$ and the covariate shift $w(x) = \mathbb{P}(C \geq c_0)/c(x)$, as shown in (6).

Operationally, the 'split version' of weighted conformal inference sketched in Algorithm 1 has three main steps:

1. split the data into a training and a calibration fold;

2. apply any prediction algorithm on the training fold to generate a *conformity score* indicating how atypical a value of the outcome is given observed covariate values;[3]

3. calibrate the predicted outcome by the distribution of conformity scores on the calibration fold. In the calibration step from Algorithm 1, Quantile$(1-\alpha; Q)$ is the $(1-\alpha)$ quantile of the distribution Q defined as

$$\text{Quantile}(1-\alpha; Q) = \sup\{z : Q(Z \leq z) < 1 - \alpha\}.$$

[3] Here, we generate a conformity score such that a large value indicates a lack of conformity to training data.

Algorithm 1: (one-sided) weighted split conformal inference

```
Input: level $\alpha$; data $\mathcal{Z} = (X_i, Y_i)_{i \in \mathcal{I}}$; testing point $x$;
       function $V(x, y; \mathcal{D})$ to compute the conformity score between $(x, y)$ and data $\mathcal{D}$;
       function $\hat{w}(x; \mathcal{D})$ to fit the weight function at $x$ using $\mathcal{D}$ as data.

Procedure:
  1. Split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{\text{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{tr}}}$ and a calibration fold $\mathcal{Z}_{\text{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{ca}}}$.
  2. For each $i \in \mathcal{I}_{\text{ca}}$, compute the conformity score $V_i = V(X_i, Y_i; \mathcal{Z}_{\text{tr}})$.
  3. For each $i \in \mathcal{I}_{\text{ca}}$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{\text{tr}}) \in [0, \infty)$.
  4. Compute the weights $\hat{p}_i(x) = \frac{W_i}{\sum_{i \in \mathcal{I}_{\text{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\text{tr}})}$ and $\hat{p}_\infty(x) = \frac{\hat{w}(x; \mathcal{Z}_{\text{tr}})}{\sum_{i \in \mathcal{I}_{\text{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\text{tr}})}$.
  5. Compute $\eta(x) = \text{Quantile}\left(1 - \alpha; \sum_{i \in \mathcal{I}_{\text{ca}}} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty\right)$.

Output: $\hat{L}(x) = \inf\{y : V(x, y; \mathcal{Z}_{\text{tr}}) \le \eta(x)\}$
```

A few comments regarding Algorithm 1 are in order. First, when the covariate shift $w(x)$ is unknown, it can be estimated using the training fold. Second, note that in step 4, if $\hat{w}(x; \mathcal{Z}_{\text{tr}}) = \infty$, then $\hat{p}_i(x) = 0$ ($i \in \mathcal{Z}_{\text{ca}}$) and $\hat{p}_\infty(x) = 1$. In this case, step 5 gives $\hat{L}(x) = -\infty$. Third, the requirement that $W_i \in [0, \infty)$ is natural because $X_i \sim P_X$ and $w(X) \in [0, \infty)$ almost surely under P_X even if Q_X is not absolutely continuous with respect to P_X. Fourth, it is worth mentioning in passing that $\eta(x)$ is invariant to positive rescalings of $\hat{w}(x)$. Thus, we can set $w(x) = 1/\hat{c}(x; c_0)$ in our case where $\hat{c}(x; c_0)$ is an estimate of $c(x; c_0)$.

In the algorithm, the conformity score function $V(x, y; \mathcal{D})$ can be arbitrary and we discuss three popular choices from the literature:

- Conformalized mean regression (CMR) scores are defined via $V(x, y; \mathcal{Z}_{\text{tr}}) = \hat{m}(x) - y$, where $\hat{m}(\cdot)$ is an estimate of the conditional mean of Y given X. The resulting LPB is then $\hat{m}(x) - \eta(x)$. This is the one-sided version of the conformity score used in Vovk et al. (2005) and Lei and Wasserman (2014).

- Conformalized quantile regression (CQR) scores are defined via $V(x, y; \mathcal{Z}_{\text{tr}}) = \hat{q}_\alpha(x) - y$, where $\hat{q}_\alpha(\cdot)$ is an estimate of the conditional α-th quantile of Y given X. The resulting LPB is then $\hat{q}_\alpha(x) - \eta(x)$. This score was proposed by Romano et al. (2019b); it is more adaptive than CMR and usually has better conditional coverage.

- Conformalized distribution regression (CDR) scores are defined via $V(x, y; \mathcal{Z}_{\text{tr}}) = \alpha - \hat{F}_{Y|X=x}(y)$, where $\hat{F}_{Y|X=x}(\cdot)$ is an estimate of the conditional distribution of Y given X. The resulting LPB is then $\hat{F}^{-1}_{Y|X=x}(\alpha - \eta(x))$, or equivalently, the $(\alpha - \eta(x))$-th quantile of the estimated conditional distribution. This score was proposed by Chernozhukov et al. (2019). It is particularly suitable to our problem because most survival analysis methods estimate the whole conditional distribution.

Under the completely independent censoring assumption, $\mathbb{P}(C \ge c_0 \mid X) = \mathbb{P}(C \ge c_0)$ almost surely. As a consequence, we can set $\hat{w}(x) = w(x) \equiv 1$ and obtain a calibrated LPB without any distributional assumption.

Proposition 1. *[Corollary 1 of Tibshirani et al. (2019)] Let c_0 be any threshold independent of $\mathcal{Z}_{\text{ca}}$. Consider Algorithm 1 with $Y_i = T_i \wedge c_0$ and $\hat{w}(x; \mathcal{D}) \equiv 1$. Under the completely independent censoring assumption, $\hat{L}(X)$ is calibrated.*

3.2 Doubly robust lower predictive bounds

Under the more general conditionally independent censoring assumption, the censoring mechanism needs to be estimated. We can apply any distributional regression techniques such as the kernel method or the newly invented distribution boosting (Friedman, 2020) to estimate $c(x) = \mathbb{P}(C \geq c_0 \mid X = x)$. For two-sided weighted split-CQR, Lei and Candès (2020) prove that the intervals satisfy a doubly robust property which states the following: the average coverage is guaranteed if either the covariate shift or the conditional quantiles are estimated well, and the conditional coverage is approximately controlled if the latter is true. In Appendix B, we generalize their results to a broad class of conformity scores proposed by Gupta et al. (2019), including the CMR-, CQR- and CDR-based scores.

In this section, we first present a version tailored to the CQR-LPB for simplicity.

Theorem 2. *Let $N = |\mathcal{Z}_{\mathrm{tr}}|, n = |\mathcal{Z}_{\mathrm{ca}}|$, c_0 be any threshold independent of $\mathcal{Z}_{\mathrm{ca}}$, and $q_\alpha(x; c_0)$ denote the α-th conditional quantile of $T \wedge c_0$ given $X = x$. Further, let $\hat{c}(x)$ and $\hat{q}(x; c_0)$ be estimates of $c(x)$ and $q(x; c_0)$ respectively using $\mathcal{Z}_{\mathrm{tr}}$, and $\hat{L}(x)$ be the corresponding CQR-LPB. Assume that there exists $\delta > 0$ such that $\mathbb{E}\left[1/\hat{c}(X)^{1+\delta}\right] < \infty$ and $\mathbb{E}[1/c(X)^{1+\delta}] < \infty$. Suppose that either A1 or A2 (or both) holds:*

A1 $\lim_{N\to\infty} \mathbb{E}\,|1/\hat{c}(X) - 1/c(X)| = 0.$

A2 *(i) There exists $b_2 > b_1 > 0$ and $r > 0$ such that, for any x and $\varepsilon \in [0, r]$,*

$$\mathbb{P}(T \wedge c_0 \geq q_\alpha(x; c_0) + \varepsilon \mid X = x) \in [1 - \alpha - b_2\varepsilon, 1 - \alpha - b_1\varepsilon]$$

(ii) $\lim_{N\to\infty} \mathbb{E}\left[\mathcal{E}(X)/\hat{c}(X)\right] = \lim_{N\to\infty} \mathbb{E}\left[\mathcal{E}(X)/c(X)\right] = 0$, *where* $\mathcal{E}(x) = |\hat{q}_\alpha(x; c_0) - q_\alpha(x; c_0)|$.

Then

$$\lim_{N,n\to\infty} \mathbb{P}\left(T \wedge c_0 \geq \hat{L}(X)\right) \geq 1 - \alpha.$$

Furthermore, under A2, for any $\varepsilon > 0$,

$$\lim_{N,n\to\infty} \mathbb{P}\Big(\mathbb{E}\big[\mathbf{1}\big\{T \wedge c_0 \geq \hat{L}(X)\big\} \mid X\big] > 1 - \alpha - \varepsilon\Big) = 1.$$

Intuitively, if $\hat{c}(x) \approx c(x)$, then the procedure approximates the oracle version of weighted split-CQR with the true weights, and the LPBs should be approximately calibrated. On the other hand, if $\hat{q}_\alpha(x; c_0) \approx q_\alpha(x; c_0)$, then $V_i \approx q_\alpha(X_i; c_0) - T_i \wedge c_0$. As a result,

$$\mathbb{P}(V_i \leq 0 \mid X_i) \approx \mathbb{P}(T_i \wedge c_0 \leq q_\alpha(X_i; c_0) \mid X_i) = \alpha.$$

Thus, the $(1-\alpha)$-th quantile of the V_i's conditional on $\mathcal{Z}_{\mathrm{tr}}$ is approximately 0. To keep on going, recall that $\eta(x)$ is the $(1-\alpha)$-th quantile of the random distribution $\sum_{i\in\mathcal{Z}_{\mathrm{ca}}} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty$, and set G to be the cumulative distribution function of this random distribution. Then,

$$G(0) \approx \mathbb{E}[G(0) \mid \mathcal{Z}_{\mathrm{tr}}] = \sum_{i\in\mathcal{Z}_{\mathrm{ca}}} \hat{p}_i(x)\mathbb{P}(V_i \leq 0 \mid \mathcal{Z}_{\mathrm{tr}}) \approx \sum_{i\in\mathcal{Z}_{\mathrm{ca}}} \hat{p}_i(x)(1-\alpha) \approx 1 - \alpha,$$

implying that $\eta(x) \approx 0$. Therefore, $\hat{L}(x) \approx q_\alpha(x; c_0)$, which approximately achieves the desired conditional coverage.

With the same intuition, we can establish a similar result for the CDR-LPB with a slightly more complicated version of Assumption **A**2.

Theorem 3. *Let $F(\cdot \mid x)$ denote the conditional distribution of $T \wedge c_0$ given $X = x$. With the same settings and assumptions as in Theorem 2, the same conclusions hold if **A**2 is replaced by the following conditions:*

(i) there exists $r > 0$ *such that, for any* x *and* $\varepsilon \in [0, r]$,

$$\mathbb{P}(T \wedge c_0 \geq q_{\alpha+\varepsilon}(x; c_0) \mid X = x) = 1 - \alpha - \varepsilon.$$

(ii) $\lim_{N\to\infty} \mathbb{E}\left[\mathcal{E}(X)/\hat{c}(X)\right] = \lim_{N\to\infty} \mathbb{E}\left[\mathcal{E}(X)/c(X)\right] = 0$, *where*

$$\mathcal{E}(x) = \sup_{s\in[\alpha,\alpha+r]} |F(\hat{q}_s(x; c_0) \mid x) - F(q_s(x; c_0) \mid x)|.$$

The double robustness of weighted split conformal inference has some appeal; indeed, the researcher can leverage knowledge about both the conditional survival function and the censoring mechanism without any concern for which is more accurate. Suppose the Cox model is adequate in a randomized clinical trial; then it can be used to produce $\hat{q}_\alpha(x; c_0)$ in conjunction with the known censoring mechanism. If the model is indeed correctly specified, the LPB is conditionally calibrated, as are classical predictive intervals derived from the Cox model (Kalbfleisch and Prentice, 2011); if the model is misspecified, however, the LPB is still calibrated.

3.3 Choice of threshold

The threshold c_0 induces an estimation-censoring tradeoff: a larger c_0 mitigates the censoring effect, closing the gap between the target outcome T and the operating outcome $T \wedge c_0$, but reduces the sample size to estimate the censoring mechanism and the conditional survival function. It is thus important to pinpoint the optimal value of c_0 to maximize efficiency.

To avoid double-dipping, we choose c_0 on the training fold $\mathcal{Z}_{\text{tr}}$. In this way, c_0 is independent of the calibration fold $\mathcal{Z}_{\text{ca}}$ and we are not using the same data twice. In particular, Proposition 1, Theorem 2 and 3 all apply. Concretely, we (1) set a grid of values for c_0, (2) randomly sample a holdout set from $\mathcal{Z}_{\text{tr}}$, (3) apply Algorithm 1 on the rest of $\mathcal{Z}_{\text{tr}}$ for each value of c_0 to generate LPBs for each unit in the holdout set, and (4) select c_0 which maximizes the average LPBs on the holdout set. One way to see all of this is to pretend that the training fold is the whole dataset and measure efficiency as the average realized LPBs. In practice, we choose 25% units from $\mathcal{Z}_{\text{tr}}$ as the holdout set. The procedure is convenient to implement, though it is by no means the most powerful approach. We leave the investigation of more principled selection procedures to future research.

4 Simulation studies

In this section, we design simulation studies to evaluate the performance of our method. Specifically, we run four sets of experiments detailed in Table 1. In each experiment, we compare the CQR- and CDR-LPB with the following alternatives:

- Cox model: we generate the LPB as the α-th quantile from an estimated Cox model. The method is implemented via the survival R-package (Therneau, 2020).
- Accelerated failure time (AFT) model: we generate the LPB as the α-th quantile from an estimated AFT model with Weibull noise. The method is implemented in the survival R package.
- Censored quantile regression: we consider three variants of quantile regression methods, proposed by Powell (1986), Portnoy (2003), and Peng and Huang (2008), respectively. All three procedures are implemented in the quantreg R package (Koenker, 2020).
- Censored quantile regression forest (Li and Bradic, 2020): this is a variant of quantile random forest (Athey et al., 2019) designed to handle time-to-event outcomes. We reimplement the method based on the code provided in https://github.com/AlexanderYogurt/censored_ExtremelyRandomForest.

- Naive CQR: we apply split-CQR (Romano et al., 2019b) naively to $(X_i, \widetilde{T}_i)_{i=1}^n$, where the quantiles are estimated by the quantreg R package.

For the CQR-LPB, the conditional quantiles are estimated via censored quantile regression forest or distribution boosting (Friedman, 2020); for the CDR-LPB, the conditional survival function is estimated via distribution boosting, which is implemented in the R package conTree (Friedman and Narasimhan, 2020).

In each experiment, we generate 200 independent datasets, each containing a training set of size $n = 3000$, and a test set of size $n = 3000$. For conformal methods, 50% of the data is used for fitting the predictive model, and the remaining 50% is reserved for calibration.[4] We then evaluate the coverage of LPBs as $(1/n_{\text{test}})\sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{T_i \geq \hat{L}(X_i)\}$. All the results in this section can be replicated with the code available at https://github.com/zhimeir/cfsurv_paper. In addition, the proposed CQR- and CDR-LPB are implemented in the R package cfsurvival, available at https://github.com/zhimeir/cfsurvival.

The covariate vector $X \in \mathbb{R}^p$ is generated from P_X. The survival time T is generated from an AFT model with Gaussian noise, i.e.

$$\log T \mid X \sim \mathcal{N}\Big(\mu(X), \sigma^2(X)\Big).$$

We consider 2×2 settings with univariate or multivariate covariates plus homoscedastic or heteroscedastic errors.[5] The choice of the parameters in each setting is specified in Table 1.

Finally, we apply all the methods with target coverage level $1 - \alpha = 90\%$. In each experiment, we estimate $c(x)$ by distribution boosting.

	dimension p	P_X	$P_{C\mid X}$	$\mu(x)$	$\sigma(x)$
Uvt. + Homosc.	1	$\mathcal{U}(0,4)$	$\mathcal{E}(0.4)$	$2 + 0.37\sqrt{x}$	1.5
Uvt. + Heterosc.	1	$\mathcal{U}(0,4)$	$\mathcal{E}(0.4)$	$2 + 0.37\sqrt{x}$	$1 + x/5$
Mvt. + Homosc.	100	$\mathcal{U}([-1,1]^p)$	$\mathcal{E}(0.4)$	$\log 2 + 1 + 0.55(x_1^2 - x_3x_5)$	1
Mvt. + Heterosc.	100	$\mathcal{U}([-1,1]^p)$	$\mathcal{E}(0.4)$	$\log 2 + 1 + 0.55(x_1^2 - x_3x_5)$	$\lvert x_{10}\rvert + 1$

Table 1: Parameters used in the simulation study. "Homosc." and "Heterosc." are short for homoscedastic and heteroscedastic; "Uvt." and "Mvt." are short for univariate and multivariate. $\mathcal{U}(a,b)$ denotes the uniform distribution supported on $[a,b]$; $\mathcal{E}(\lambda)$ denotes the exponential distribution with rate λ.

Figure 1 presents the empirical coverage of the LPBs on uncensored survival times. Censored random forests, the Cox model, the AFT model, and the three quantile regression methods fail to achieve the target coverage in most cases. On the other hand, the naive CQR attains the desired coverage but at the price of being overly conservative. In contrast, both the CQR- and CDR-LPB achieve near-exact marginal coverage, as predicted by our theory.

Next, we investigate the conditional coverage and efficiency of these methods. In Figure 2(a), we plot the empirical conditional coverage as a function of the conditional variance of T on X. In particular, we stratify the data into 10 groups based on equispaced percentiles of $\text{Var}(T \mid X)$ and plot the average coverage within each stratum along with a 90% confidence band obtained via repeated sampling. (Note that in either the homoscedastic or the heteroscedastic case, $\text{Var}(T \mid X)$ is varying with X.) Not surprisingly, the naive CQR is conditionally conservative. In the univariate case, both the CQR- and CDR-LPB approximately achieve desired conditional coverage; in the

[4] The splitting ratio between the training set and the test set is slightly different from the recommendation from Sesia and Candès (2020), where they suggest using 75% of the data for training and 25% for calibration. We reserve more data for calibration to ensure there are still enough samples in the calibration set after the selection and to decrease the variability of the LPBs.

[5] Here the term "homoscedastic" or "heteroscedastic" is applied to $\log T$.

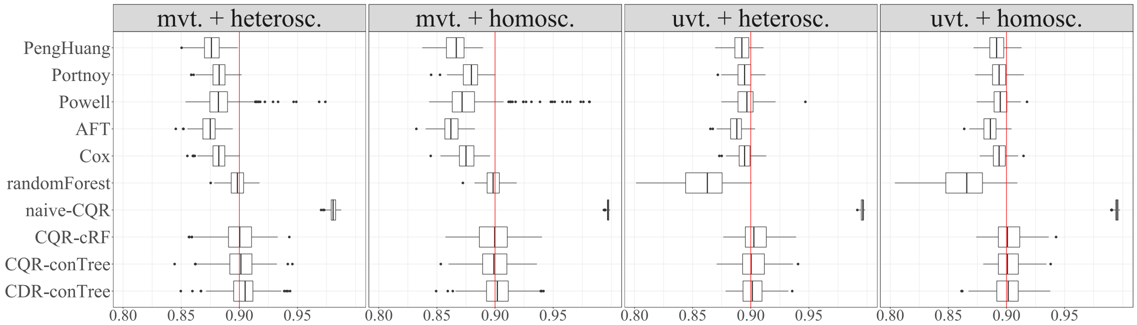


Figure 1: Empirical 90% coverage of the uncensored survival time T. "CQR-cRF" is short for the CQR-LPB with censored quantile regression forest; "CQR-conTree" and "CDR-conTree" are short for the CQR- and CDR-LPB with distribution boosting. The other abbreviations are the same as in Table 1.

multivariate case, the conditional coverage is slightly uneven, though still concentrating around the target line. Figure 2(b) presents the ratio between the LPBs and the true α-th conditional quantile as a function of $\mathrm{Var}(T \mid X)$. This is a measure of efficiency since the true conditional quantile is the oracle LPB. Here, we observe that naive CQR-LPBs are close to zero, confirming that they are overly conservative, while the CQR- and CDR-LPBs are fairly close to the oracle LPB, implying that both methods are relatively efficient.

5 Application to UK Biobank COVID-19 data

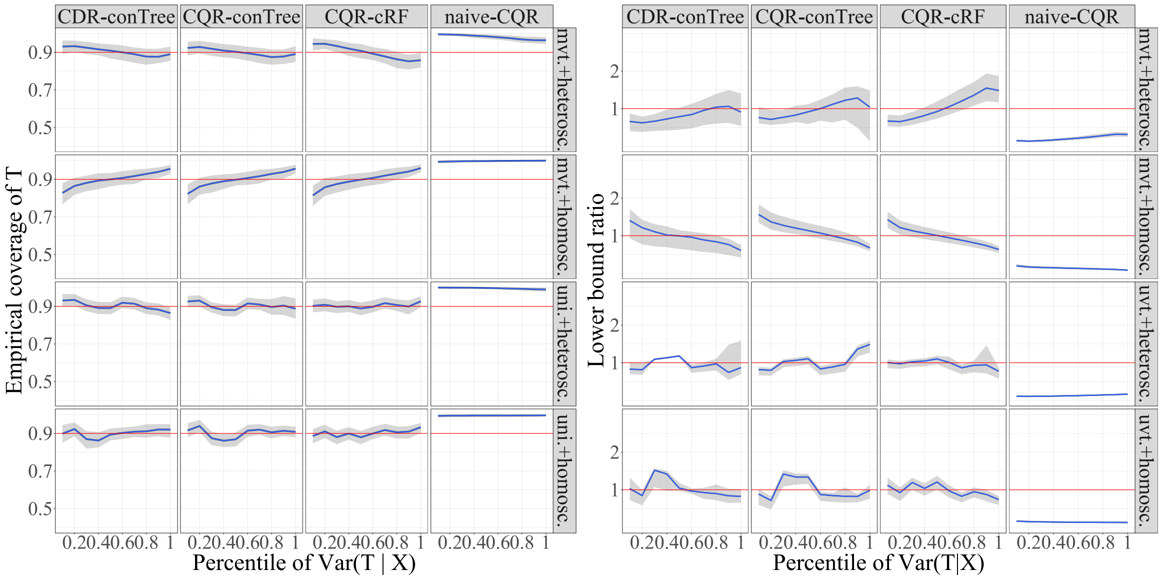
We apply our method to the UK Biobank COVID-19 dataset to demonstrate robustness and practicability. UK Biobank (Bycroft et al., 2018) is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. In April 2020, UK Biobank started to release COVID-19 testing data, and has since continued to regularly provide updates. This gives researchers access to a cohort of COVID-19 patients, along with their date of confirmation, survival status, pre-existing conditions, and other demographic covariates.

We include in our analysis all individuals in UK Biobank who received a positive COVID-19 test result before January 21st, 2021. This results in a dataset of size $n = 14{,}861$ with 484 events, defined as a COVID-related death. We extract eight covariate features, namely, *age, gender, body mass index (BMI), waist size, cardiovascular disease status, diabetes status, hypothyroidism status, and respiratory disease status.* As in Section 2, the censoring time is the time lapse between the date of a positive test and January 21st, 2021. The survival time is the time lapse between the date of a positive test and the event (which may have yet to occur).

We wish to harness this data to produce an LPB on the survival time of each COVID-19 patient. To apply the CQR- or CDR-LPB, we set the threshold c_0 to be 14 days. Since survival time assessment likely informs high-stakes decision-making, we set the target level to 99% for reliability.

5.1 Semi-synthetic examples

To demonstrate robustness, we start our analysis with two semi-synthetic examples so that the ground truth is known and calibration can be assessed (results on real outcomes are presented next). We keep the covariate matrix X from the UK Biobank COVID-19 data. In the first simulation study, we substitute the censoring time with a synthetic C. In the second, each survival time, observed or not, is substituted with a synthetic version. Details follow:



(a)

(b)

Figure 2: Results from the experiments detailed in Table 1: (a) empirical 90% conditional coverage and (b) ratio between the LPB and the theoretical quantile as a function of $\text{Var}(T \mid X)$. The blue curves correspond to the mean coverage in (a) and the median ratio in (b). The gray confidence bands correspond to the 95% and 5% quantiles of the estimates over repeated sampling. The abbreviations are the same as in Figure 1.

- *Synthetic* C: we take the censored survival time $\widetilde{T}$ as the uncensored survival time and generate the censoring time C_{syn} as

  $$C_{\text{syn}} \sim \mathcal{E}(0.001 \cdot \text{age} + 0.01 \cdot \text{gender}).$$

  In this setting, the observables are $(X, C_{\text{syn}}, \widetilde{T} \wedge C_{\text{syn}})$, and we wish to construct LPBs on $\widetilde{T}$.

- *Synthetic* T: we keep the real censoring time C, and generate a survival time T_{syn} as:

  $$\log T_{\text{syn}} \mid X \sim \mathcal{N}(2 + 0.05 \cdot \text{age} + 0.1 \cdot \text{gender}, 1).$$

  In this setting, the observables are $(X, C, T_{\text{syn}} \wedge C)$, and we wish to construct LPBs on T_{syn}.

Figure 3 shows the histograms of the survival time, censoring time, and censored survival time from the two simulated datasets. We apply the CDR-LPB (with $c_0 = 14$) to both. For comparison, we also apply the AFT and naive CQR. To evaluate the LPBs, we randomly split the data into a training set with 75% of the data and a holdout set with the remaining 25%. Each method is applied to the training set, and the resulting LPBs are evaluated on the holdout set. We repeat the above procedure 100 times to create 100 pairs of training and test data sets.

To visualize conditional calibration, we fit a Cox model on the data to generate a predicted risk score for each unit and stratify all units into 10 subgroups defined by deciles of the predicted risk. The results for synthetic C and T are plotted in Figures 4 and 5, respectively. As in the simulation studies from Section 4, we see that the naive CQR is overly conservative. Notably, although the AFT-LPB is well calibrated in the synthetic-C setting, this method is overly conservative in the synthetic-T setting, even though the model is correctly specified. In contrast, the CDR-LPB is calibrated in both examples. From the middle panels of Figures 4 and 5, we also observe that the

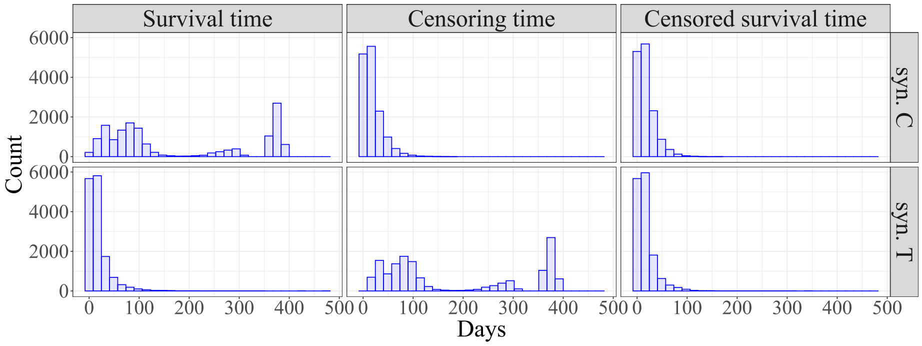


Figure 3: Histograms of the survival time, censoring time, and censored survival time defined as the minimum between the two, in each simulation setting.

CDR-LPB is approximately conditionally calibrated. Finally, the right panels show that CDR-LPB nearly preserves the rank of the predicted risk given by the Cox model. The flat portion of the LPB towards the left end corresponds to the threshold, implying that at least 99% of people with predicted risk scores lower than 0.5 can survive beyond 14 days.

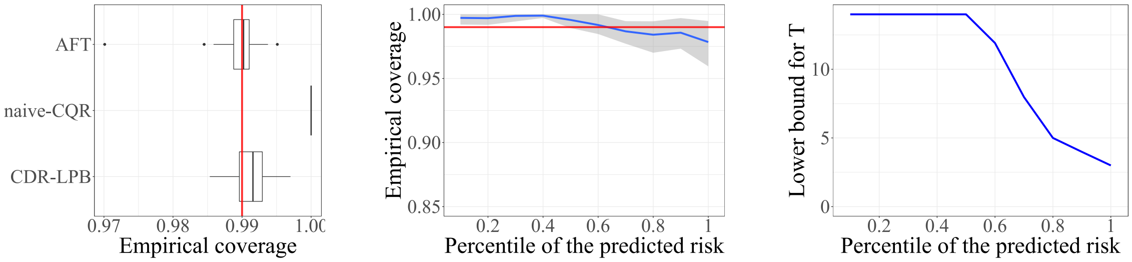


Figure 4: Results for synthetic censoring times across 100 replications: empirical coverage (left), empirical conditional coverage of the CDR-LPB (middle), and CDR-LPB (right) as a function of the percentile of the predicted risk. The target coverage level is 99%. The blue curves correspond to the mean coverage in the middle panel and the median LPB in the right panel; the gray confidence bands correspond to the 5% and 95% quantiles of the estimates across 100 independent replications.

5.2 Real data analysis

We now turn attention to actual COVID-19 responses. Again, we randomly split the data into a training set including 75% of data and a holdout set including the remaining 25%. Then we run the CDR on the training set and validate the LPBs on the holdout set. The issue is that the actual survival time is only partially observed, and thus, the coverage of a given LPB cannot be assessed accurately (this is precisely why we generated semi-synthetic responses in the previous section.) Nevertheless, we note that

$$\beta_{\mathrm{lo}} := \mathbb{P}\big(\widetilde{T} \geq \hat{L}(X)\big) \leq \mathbb{P}\big(T \geq \hat{L}(X)\big) \leq 1 - \mathbb{P}\big(\widetilde{T} < \hat{L}(X), T \leq C\big) =: \beta_{\mathrm{hi}},$$

where both β_{lo} and β_{hi} are estimable from the data. This says that we can assess the marginal coverage of the LPBs by evaluating a lower and upper bound on the coverage. Of course, this extends to conditional coverage.

To assess the stability, we evaluate our method on 100 independent sample splits. Figure 6 presents the empirical lower and upper bound of the marginal coverage and those of the conditional

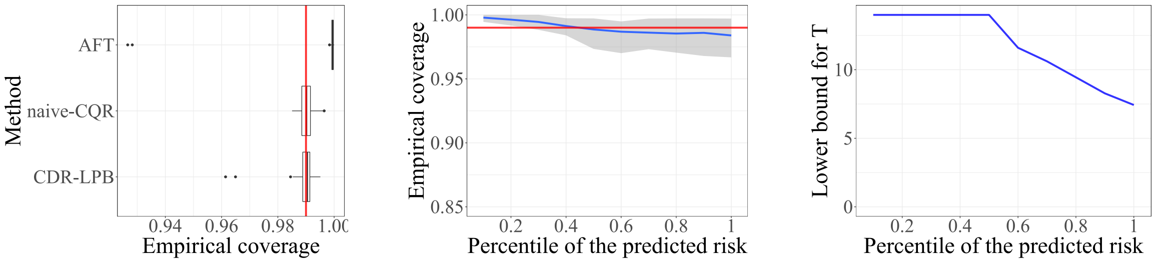


Figure 5: Results for synthetic survival times: everything else is as in Figure 4.

coverage as functions of the predicted risk (as in the semi-synthetic examples), together with their variability across 100 sample splits. The left panel shows that the upper bound is very close to the lower bound, and both concentrate around the target level. Thus we can be assured that the CDR-LPB is well calibrated. Similarly, the other panels show that the CDR-LPB is approximately conditionally calibrated. We conclude this section by showing in Figure 7 the LPBs as functions of the percentiles of the predicted risk, age, and BMI, respectively.

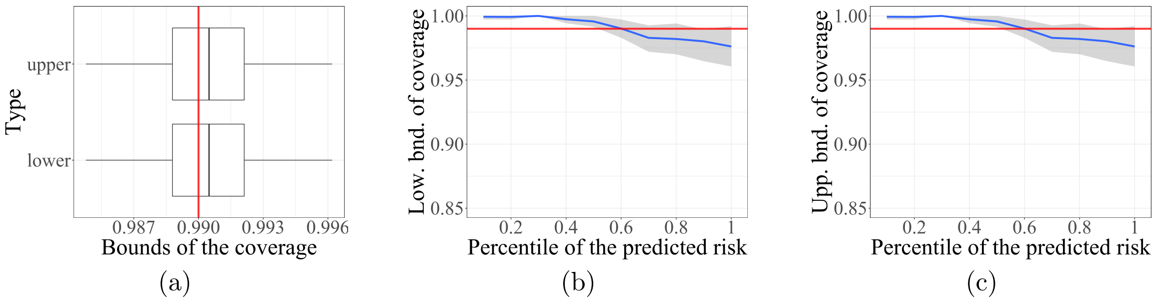


Figure 6: Analysis of the UK Biobank COVID-19 data: (a) lower and upper bounds of the empirical coverage; (b) lower and (c) upper bounds of empirical coverage as a function of the predicted risk. The target coverage level is 99%. The blue curves correspond to the mean coverage, and the gray confidence bands correspond to the 5% and 95% quantiles of the estimates across 100 sample splits.

6 Discussion and extensions

6.1 Sharper coverage criteria

It is more desirable to achieve a stronger conditional coverage criterion:

$$\mathbb{P}\left(T \geq \hat{L}(X) \mid X = x\right) \geq 1 - \alpha, \tag{8}$$

which states that $\hat{L}(X)$ is a conditionally calibrated LPB. Clearly, (8) implies valid marginal coverage. Theorem 2 and 3 show that the CQR- and CDR-LPB are approximately conditionally calibrated if the conditional quantiles are estimated well. However, without distributional assumptions, we can show that (8) can only be achieved by trivial LPBs.

Theorem 4. *Assume that $X \in \mathbb{R}^p$ and $C \geq 0, T \geq 0$. Let $P_{(X,C)}$ be any given distribution of (X, C). If $\hat{L}(\cdot)$ satisfies (8) uniformly for all joint distributions of (X, C, T) with $(X, C) \sim P_{(X,C)}$, then for all such distributions,*

$$\mathbb{P}(\hat{L}(x) = 0) \geq 1 - \alpha,$$

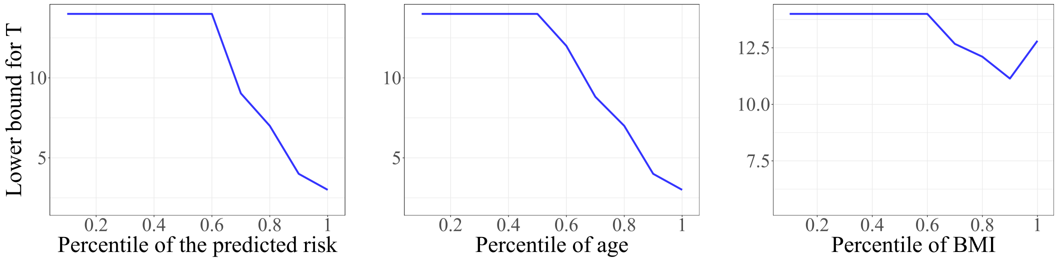


Figure 7: Analysis of the UK Biobank COVID-19 data: LPBs on the survival time of COVID-19 patients as a function of the percentiles of predicted risk (left), age (middle) and BMI (right). The target coverage level is 99%. The blue curves correspond to the median LPB across 100 sample splits.

at almost surely all points x aside from the atoms of P_X.

Theorem 4 implies that no nontrivial LPB exists even if the distribution of (X, C) is known. Put another way, it is impossible to achieve desired conditional coverage while being agnostic to the conditional survival function. This impossibility result is inspired by previous works on uncensored outcomes and two-sided intervals (Vovk, 2012; Barber et al., 2019a).

It is valuable to find other achievable coverage criteria which are sharper than the marginal coverage criterion (3). Without censoring and covariate shift, Vovk et al. (2003) introduced Mondrian conformal inference to achieve desired marginal coverage over multiple subpopulations. The idea is further developed from different perspectives (Vovk, 2012; Lei et al., 2013; Guan, 2019; Barber et al., 2019a; Romano et al., 2019a). Given a partition of the covariate space $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$, Mondrian conformal inference guarantees that[6]

$$\mathbb{P}(Y \in \hat{C}(X) \mid X \in \mathcal{X}_k) \geq 1 - \alpha, \quad k = 1, \ldots, K.$$

Following their techniques, we can extend Mondrian conformal inference to our case by modifying the calibration term $\eta(x)$ (in Algorithm 1):

$$\eta(x) = \text{Quantile}\left(1 - \alpha; \sum_{i \in \mathcal{I}_{\text{ca}}, X_i \in \mathcal{X}_k} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty\right), \quad \forall x \in \mathcal{X}_k. \tag{9}$$

Suppose $\mathcal{X}_1$ and $\mathcal{X}_2$ correspond to male and female subpopulations. Then $\eta(x)$ is a function of both the testing point x and the gender. That said, estimation of censoring mechanisms and conditional survival functions can still depend on the whole training fold $\mathcal{Z}_{\text{tr}}$ as joint training may be more powerful than separate training on each subpopulation (Romano et al., 2019a).

When the censoring mechanism is known, we can prove that

$$\mathbb{P}(T \wedge c_0 \geq \hat{L}(X) \mid X \in \mathcal{X}_k) \geq 1 - \alpha, \quad k = 1, \ldots, K. \tag{10}$$

By the conditionally independent censoring assumption, the target distribution in the localized criterion (10) for a given k can be rewritten as

$$(X, T \wedge c_0) \mid C \geq c_0, X \in \mathcal{X}_k \;\sim\; P_{X \mid C \geq c_0, X \in \mathcal{X}_k} \times P_{T \wedge c_0 \mid X}.$$

[6] Mondrian conformal inference allows the subgroups to also depend on the outcome; see Vovk et al. (2005), which refers to the rule of forming subgroups as a “taxonomy.” Besides, the subgroups can also be overlapping; see Barber et al. (2019a).

The covariate shift between the observed and target distributions is

$$w_k(x) = \frac{dP_{X \mid C \geq c_0, X \in \mathcal{X}_k}}{dP_X}(x) \propto \frac{I(x \in \mathcal{X}_k)}{\mathbb{P}(C \geq c_0 \mid X = x)}.$$

This justifies the calibration term (9) in the weighted Mondrian conformal inference. Since the weighted Mondrian conformal inference is a special case of Algorithm 1, it also enjoys the double robustness property, implied by Theorem B.1 in Appendix B.

6.2 Survival counterfactual prediction

The proposed method in this paper is designed for a single cohort. In practice, patients are often exposed to multiple conditions, and the goal is to predict the counterfactual survival times had the cohort been exposed to a different condition. For example, a clinical study typically involves a treatment group and a control group. For a new patient, it is of interest to predict her survival time had she been assigned the treatment. For uncensored outcomes, Lei and Candès (2020) proposed a method based on weighted conformal inference for counterfactual prediction under the potential outcome framework (Neyman, 1990; Rubin, 1974). We can extend their strategy to handle censored outcomes and apply it to the survival counterfactual prediction.

Suppose each patient has a pair of potential survival times $(T(1), T(0))$, where $T(1)$ (resp. $T(0)$) denotes the survival time had the patient been assigned into the treatment (resp. control) group. Our goal is to construct a calibrated LPB on $T(1)$, given i.i.d. observations $(X_i, W_i, C_i, T_i)_{i=1}^n$ with W_i denoting the treatment assignment and

$$T_i = \begin{cases} T_i(1), & W_i = 1, \\ T_i(0), & W_i = 0. \end{cases}$$

Without further assumptions on the correlation structures between $T(1)$ and $T(0)$, it is natural to conduct inference based on the observed treated group since the control group contains no information about $T(1)$. The joint distribution of $(X, T(1) \wedge c_0)$ on this group becomes

$$(X, T(1) \wedge c_0) \mid C \geq c_0, W = 1 \;\sim\; P_{X \mid C \geq c_0, W=1} \times P_{T(1) \wedge c_0 \mid X, C \geq c_0, W=1}.$$

Under the assumption that $(T(1), T(0)) \perp\!\!\!\perp (W, C) \mid X$, the conditional distribution of $T(1) \wedge c_0$ matches the target:

$$P_{T(1) \wedge c_0 \mid X, C \geq c_0, W=1} = P_{T(1) \wedge c_0 \mid X}.$$

The assumption is a combination of the strong ignorability assumption (Rubin, 1978), a widely accepted starting point in causal inference, and the conditionally independent censoring assumption. The density ratio of the two covariate distributions can be characterized by

$$w(x) = \frac{dP_{X \mid C \geq c_0, W=1}}{dP_X}(x) \propto \frac{1}{\mathbb{P}(C \geq c_0, W = 1 \mid X = x)}.$$

In many applications, it is plausible to further assume that $C \perp\!\!\!\perp W \mid X$. In this case,

$$\mathbb{P}(C \geq c_0, W = 1 \mid X = x) = \mathbb{P}(C \geq c_0 \mid X = x)\mathbb{P}(W = 1 \mid X = x),$$

where the first term is the censoring mechanism and the second term is the propensity score (Rosenbaum and Rubin, 1983). Therefore, we can obtain calibrated LPBs on counterfactual survival times if both the censoring mechanism and the propensity score are known. This assumption is often plausible for randomized clinical trials. Furthermore, it has a doubly robust guarantee of coverage that is similar to Theorems 2 and 3.

Acknowledgment

E. C. was supported by Office of Naval Research grant N00014-20-12157, by the National Science Foundation grants OAC 1934578 and DMS 2032014, by the Army Research Office (ARO) under grant W911NF-17-1-0304, and by the Simons Foundation under award 814641. L. L. and Z. R. were supported by the same NSF OAC grant and by the Discovery Innovation Fund for Biomedical Data Sciences. L.L. was also supported by NIH grant R01MH113078. The authors are grateful to Ying Jin, Yan Min, Chiara Sabatti, Matteo Sesia, Lu Tian and Steve Yadlowsky for their constructive feedback.

References

Aitchison, J. and Dunsmore, I. R. (1980). *Statistical prediction analysis*. CUP Archive.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Number 46. Sage.

Anisimov, V. V. and Fedorov, V. V. (2007). Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in medicine*, 26(27):4958–4975.

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

Bain, L. (2017). *Statistical analysis of reliability and life-testing models: theory and methods*. Routledge.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019b). Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*.

Barnard, K. D., Dent, L., and Cook, A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC medical research methodology*, 10(1):1–8.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Carter, R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. *Controlled clinical trials*, 25(5):429–436.

Carter, R. E., Sonne, S. C., and Brady, K. T. (2005). Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, 5(1):1–5.

Cauchois, M., Gupta, S., and Duchi, J. (2020). Knowing what you know: valid confidence sets in multiclass and multilabel prediction. *arXiv preprint arXiv:2004.10181*.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2019). Distributional conformal prediction. *arXiv preprint arXiv:1909.07889*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.

Emanuel, E., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., and Phillips, J. (2020). Fair allocation of scarce medical resources in the time of covid-19. *The New England Journal of Medicine*, 382.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in medicine*, 14(1):73–82.

Friedman, J. and Narasimhan, B. (2020). *conTree: Contrast Trees and Boosting.* R package version 0.2-8.

Friedman, J. H. (2020). Contrast trees and distribution boosting. *Proceedings of the National Academy of Sciences*, 117(35):21175–21184.

Gajewski, B. J., Simon, S. D., and Carlson, S. E. (2008). Predicting accrual in clinical trials with bayesian posterior predictive distributions. *Statistics in medicine*, 27(13):2328–2340.

Geisser, S. (1993). *Predictive inference*, volume 55. CRC press.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical journal*, 52(1):70–84.

Guan, L. (2019). Conformal prediction with localization. *arXiv preprint arXiv:1908.08558.*

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. (2019). Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562.*

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer.

Hong, H. and Tamer, E. (2003). Inference in censored models with endogenous regressors. *Econometrica*, 71(3):905–932.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841–860.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2016). Deep survival: A deep cox proportional hazards network. *stat*, 1050(2).

Koenker, R. (2020). *quantreg: Quantile Regression*. R package version 5.75.

Krishnamoorthy, K. and Mathew, T. (2009). *Statistical tolerance regions: theory, applications, and computation*, volume 744. John Wiley & Sons.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156.

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., and Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):1–8.

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(523):1094–1111.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96.

Lei, L. and Candès, E. J. (2020). Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*.

Leung, K.-M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104.

Li, A. H. and Bradic, J. (2020). Censored quantile regression forest. In *International Conference on Artificial Intelligence and Statistics*, pages 2109–2119. PMLR.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5:465–472. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1–51 (Annals of Agricultural Sciences).

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012.

Powell, J. L. (1986). Censored regression quantiles. *Journal of econometrics*, 32(1):143–155.

Ranney, M. L., Griffeth, V., and Jha, A. K. (2020). Critical supply shortages—the need for ventilators and personal protective equipment during the covid-19 pandemic. *New England Journal of Medicine*, 382(18):e41.

Ratkovic, M. and Tingley, D. (2021). Estimation and inference on nonlinear and heterogeneous effects. Technical report.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. J. (2019a). With malice towards none: Assessing uncertainty via equalized coverage. *arXiv preprint arXiv:1908.05428*.

Romano, Y., Patterson, E., and Candes, E. (2019b). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3543–3553.

Romano, Y., Sesia, M., and Candès, E. J. (2020). Classification with valid and adaptive coverage. *arXiv preprint arXiv:2006.02544*.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Sadinle, M., Lei, J., and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Sant'Anna, P. H. (2016). Program evaluation with right-censored data. *arXiv preprint arXiv:1604.02642*.

Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 722–726.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634.

Sesia, M. and Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261.

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

Shah, R. D., Peters, J., et al. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.2-7.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32:2530–2540.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Vergano, M., Bertolini, G., Giannini, A., Gristina, G., Livigni, S., Mistraletti, G., Riccioni, L., and Petrini, F. (2020). Clinical ethics recommendations for the allocation of intensive care treatments in exceptional, resource-limited circumstances: the italian perspective during the covid-19 epidemic. *Critical Care*, 24.

Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314.

von Bahr, B., Esseen, C.-G., et al. (1965). Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Vovk, V. (2002). On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196. IEEE.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V., Lindsay, D., Nouretdinov, I., and Gammerman, A. (2003). Mondrian confidence machine. *Technical Report*.

Wald, A. (1943). An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

A Proofs of impossibility results

A.1 Proof of Theorem 1

Let $(X_i, C_i, T_i)_{i=1}^{n+1} \overset{i.i.d.}{\sim} (X, C, T)$. For notational convenience, we put $Z_i = (X_i, C_i, T_i)$. To avoid confusion, we expand $\hat{L}(x)$ into $\hat{L}(Z_1, \ldots, Z_n; x)$. Note that $\hat{L}$ depends on $(Z_1, \ldots, Z_n)$ through $(X_i, C_i, \widetilde{T}_i)_{i=1}^n$. Let

$$\phi(Z_1, \ldots, Z_{n+1}) = I(T_{n+1} < \hat{L}_n(Z_1, \ldots, Z_n; X_{n+1})).$$

Since $\hat{L}_n$ satisfies (3) under the conditionally independent censoring assumption (1), we have that

$$\mathbb{P}\left(\phi(Z_1, \ldots, Z_{n+1}) = 1\right) \leq \alpha.$$

As a result, if we treat $T \perp\!\!\!\perp C \mid X$ as a null hypothesis, $\phi(Z_1, \ldots, Z_{n+1})$ is an α-level test. Note that X is continuous, and (T, C) are continuous or discrete. By Theorem 2 and Remark 4 of Shah et al. (2020), for any joint distribution Q of Z with the same continuity conditions on (X, C, T),

$$\mathbb{P}_{Z_i \overset{i.i.d.}{\sim} Q}(\phi(Z_1, \ldots, Z_{n+1}) = 1) \leq \alpha. \tag{A.1}$$

Let $\tilde{Z}_i = (X_i, C_i, \widetilde{T}_i)$ and Q denote its distribution. Then $\widetilde{T}_i \wedge C_i = T_i \wedge C_i$ and thus

$$\hat{L}_n(\tilde{Z}_1, \ldots, \tilde{Z}_n; x) = \hat{L}_n(Z_1, \ldots, Z_n; x).$$

Clearly, X is absolutely continuous with respect to the Lebesgue measure and $\widetilde{T}, C$ are absolutely continuous with respect to the Lebesgue measure or the counting measure. By (A.1) and the definition of ϕ, we have

$$\mathbb{P}\left(\tilde{T}_{n+1} \geq \hat{L}_n(Z_1, \ldots, Z_n; X_{n+1})\right) \geq 1 - \alpha.$$

The proof is then completed by replacing (T_{n+1}, X_{n+1}) with (T, X).

A.2 Proof of Theorem 4

We prove the theorem by modifying the proof of Proposition 4 from Vovk (2012). To avoid confusion, we expand $\hat{L}(x)$ into $\hat{L}(Z_1, \ldots, Z_n; x)$ where $Z_i = (X_i, C_i, \widetilde{T}_i)$. Fix any distribution P with $(X, C) \sim P_{(X,C)}$ and $C \geq 0, T \geq 0$ almost surely. Suppose there exists a set $\mathcal{V}$ of P_X-non-atom x such that $P_X(\mathcal{V}) > 0$, and for any $x \in \mathcal{V}$,

$$\mathbb{P}_{Z_i \overset{i.i.d.}{\sim} P}(\hat{L}(Z_1, \ldots, Z_n, x) > 0) > \alpha.$$

Since $\mathcal{V}$ only includes non-atom x's, there exists $t_0 > 0$ and $\delta > 0$ such that

$$\mathbb{P}_{Z_i \overset{i.i.d.}{\sim} P}(\hat{L}(x) > t_0) > \alpha + \delta, \quad \forall x \in \mathcal{V}. \tag{A.2}$$

We can further shrink $\mathcal{V}$ so that

$$\sqrt{2 - 2(1 - P_X(\mathcal{V}))^n} \leq \delta/2. \tag{A.3}$$

Fix any $t_1 \in (0, t_0)$. Define a new probability distribution Q on (X, C, T) with $(X, C) \sim P_{(X,C)}$ and the regular conditional probability

$$Q(T \in A \mid X = x, C = c) = \begin{cases} P(T \in A \mid X = x, C = c), & x \notin \mathcal{V}, \\ \delta_{t_1}(A), & x \in \mathcal{V}, \end{cases}$$

where δ_{t_1} defines the point mass on t_1. Let d_{TV} denote the total-variation distance. Then,

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= \sup_{\mathcal{A}_X, \mathcal{A}_C, \mathcal{A}_T} |P(X \in \mathcal{A}_X, C \in \mathcal{A}_C, T \in \mathcal{A}_T) - Q(X \in \mathcal{A}_X, C \in \mathcal{A}_C, T \in \mathcal{A}_T)| \\ &= \sup_{\mathcal{A}_X, \mathcal{A}_C, \mathcal{A}_T} |P(X \in \mathcal{A}_X \cap \mathcal{V}, C \in \mathcal{A}_C, T \in \mathcal{A}_T) - Q(X \in \mathcal{A}_X \cap \mathcal{V}, C \in \mathcal{A}_C, T \in \mathcal{A}_T)| \\ &\leq \sup_{\mathcal{A}_X, \mathcal{A}_C, \mathcal{A}_T} \max\{P(X \in \mathcal{A}_X \cap \mathcal{V}, C \in \mathcal{A}_C, T \in \mathcal{A}_T), Q(X \in \mathcal{A}_X \cap \mathcal{V}, C \in \mathcal{A}_C, T \in \mathcal{A}_T)\} \\ &\leq \sup_{\mathcal{A}_X} \max\{P(X \in \mathcal{A}_X \cap \mathcal{V}), Q(X \in \mathcal{A}_X \cap \mathcal{V})\} \\ &= \sup_{\mathcal{A}_X} P_X(\mathcal{A}_X \cap \mathcal{V}) \leq P_X(\mathcal{V}). \end{aligned}$$

Using the tensorization inequality for the total-variation distance (see e.g., Tsybakov (2008), Section 2.4) and (A.3), we obtain that

$$d_{\mathrm{TV}}(P^n, Q^n) \le \sqrt{2 - 2(1 - d_{\mathrm{TV}}(P, Q))^n} \le \delta/2.$$

Together with (A.2), this implies that

$$\mathbb{P}_{Z_i \overset{i.i.d.}{\sim} Q}(\hat{L}(x) > t_0) > \alpha + \delta/2, \quad \forall x \in \mathcal{V}.$$

Let $Z = (X, C, T)$ be an independent draw from Q. The above inequality can be reformulated as

$$\mathbb{P}_{Z_i, Z \overset{i.i.d.}{\sim} Q}(\hat{L}(X) > t_0 \mid X = x) > \alpha + \delta/2, \quad \forall x \in \mathcal{V}.$$

Marginalizing over $x \in \mathcal{V}$, it implies that

$$\mathbb{P}_{Z_i, Z \overset{i.i.d.}{\sim} Q}(\hat{L}(X) > t_0, X \in \mathcal{V}) > (\alpha + \delta/2) Q_X(\mathcal{V}). \tag{A.4}$$

By definition of Q, $T = t_1 < t_0$ almost surely conditional on $X \in \mathcal{V}$. Thus,

$$\mathbb{P}_{Z_i, Z \overset{i.i.d.}{\sim} Q}(T < \hat{L}(X), X \in \mathcal{V}) > (\alpha + \delta/2) Q_X(\mathcal{V}).$$

On the other hand, since Q is a distribution with the same marginal distribution of (X, C) and $T \ge 0$ almost surely, for any x,

$$\mathbb{P}_{Z_i, Z \overset{i.i.d.}{\sim} Q}(T < \hat{L}(X) \mid X = x) \le \alpha.$$

Marginalizing over $x \in \mathcal{V}$, it implies that

$$\mathbb{P}_{Z_i, Z \overset{i.i.d.}{\sim} Q}(T < \hat{L}(X), X \in \mathcal{V}) \le \alpha Q_X(\mathcal{V}).$$

This contradicts (A.4) since $Q_X(\mathcal{V}) = P_X(\mathcal{V}) > 0$. The theorem is proved by contradiction.

B Double robustness of weighted conformal inference

B.1 conformity score via nested sets

Gupta et al. (2019) introduced a broad class of conformity scores characterized by nested sets. Suppose we have a totally ordered index set $\mathcal{S}$ (e.g., $\mathbb{R}$) and a sequence of nested sets $\{\mathcal{F}_s(x; \mathcal{D}) : s \in \mathcal{S}\}$ in the sense that $\mathcal{F}_{s_1}(x; \mathcal{D}) \subset \mathcal{F}_{s_2}(x; \mathcal{D})$ for any $s_1 \le s_2 \in \mathcal{S}$. Define a score as the index of the minimal set that includes y, i.e.

$$V(x, y; \mathcal{D}) = \inf\{s \in \mathcal{S} : y \in \mathcal{F}_s(x; \mathcal{D})\}.$$

Without loss of generality, we assume throughout that $\mathcal{F}_{\inf \mathcal{S}}(x)$ is the empty set and $\mathcal{F}_{\sup \mathcal{S}}(x)$ is the full domain of Y. The CMR-, CQR-, and CDR-based scores are instances of this:

- CMR score: $\mathcal{F}_s(x; \mathcal{D}) = [\hat{m}(x) - s, \infty)$.
- CQR score: $\mathcal{F}_s(x; \mathcal{D}) = [\hat{q}_\alpha(x) - s, \infty)$.
- CDR score: $\mathcal{F}_s(x; \mathcal{D}) = [\hat{q}_{\alpha - s}(x), \infty)$.

We refer the readers to Table 1 of Gupta et al. (2019) for a list of other conformity scores.

B.2 A general double robustness property

Throughout the rest of this section, for any event A and random variable X we write $\mathbb{P}(A \mid X)$ for the conditional expectation $\mathbb{E}[\mathbf{1}_A \mid X]$; clearly, $\mathbb{P}(A \mid X)$ is a function of X.

Theorem B.1. *Let $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y) \sim P_X \times P_{Y|X}$ and let Q_X be another distribution on the domain of X. Set $N = |\mathcal{Z}_{\mathrm{tr}}|$ and $n = |\mathcal{Z}_{\mathrm{ca}}|$. Further, let $\{\mathcal{F}_s(x; \mathcal{D}) : s \in \mathcal{S}\}$ be any sequence of nested sets, $\hat{w}(x) = \hat{w}(x; \mathcal{Z}_{\mathrm{tr}})$ be an estimate of $w(x) = (dQ_X/dP_X)(x)$, and $\hat{C}(x)$ be the resulting conformal interval from Algorithm 1. Assume that $\mathbb{E}[\hat{w}(X) \mid \mathcal{Z}_{\mathrm{tr}}] = 1$ and $\mathbb{E}[w(X)] = 1$. Assume that either B1 or B2 (or both) holds:*

B1 $\lim_{N\to\infty} \mathbb{E}\left|\hat{w}(X) - w(X)\right| = 0$;

B2 *(a)* $\mathbb{P}_{X\sim Q_X}(w(X) < \infty) = 1$, *and there exists $\delta > 0$ such that* $\limsup_{N\to\infty} \mathbb{E}\left[\hat{w}(X)^{1+\delta}\right] < \infty$;

(b) There exists $r > 0, s_0 \in \mathcal{S}$, $b_2 > b_1 > 0$, and a sequence of oracle nested sets $\{\mathcal{O}_s(x)\}_{s\in\mathcal{S}}$, such that

(i) for any $\varepsilon \in [0, r]$,

$$1 - \alpha - b_2\varepsilon \le \mathbb{P}\left(Y \in \mathcal{O}_{s_0-\varepsilon}(X) \mid X\right) \le 1 - \alpha - b_1\varepsilon, \quad \textit{almost surely};$$

(ii) $\lim_{N\to\infty} \mathbb{E}\left[\hat{w}(X)\Delta(X)\right] = \lim_{N\to\infty} \mathbb{E}\left[w(X)\Delta(X)\right] = 0$, *where* $\Delta(x) = \sup_{s\in[s_0-r,s_0]} \Delta_s(x)$,

$$\Delta_s(x) = \inf\left\{\Delta \ge 0 : s - \Delta \in \mathcal{S}, \mathcal{F}_{s-\Delta}(x; \mathcal{Z}_{\mathrm{tr}}) \subset \mathcal{O}_s(x) \text{ and } \mathcal{O}_{s-\Delta}(x) \subset \mathcal{F}_s(x; \mathcal{Z}_{\mathrm{tr}})\right\}.$$

Then

$$\lim_{N,n\to\infty} \mathbb{P}_{(X,Y)\sim Q_X\times P_{Y|X}}\left(Y \in \hat{C}(X)\right) \ge 1 - \alpha. \tag{B.1}$$

Furthermore, under B2, for any $\varepsilon > 0$,

$$\lim_{N,n\to\infty} \mathbb{P}_{X\sim Q_X}\left(\mathbb{P}\left(Y \in \hat{C}(X) \mid X\right) \le 1 - \alpha - \varepsilon\right) = 0. \tag{B.2}$$

The proof of Theorem B.1 is a generalization of Theorem A.1 of Lei and Candès (2020). Here we present a self-contained proof for completeness. We start with three lemmas.

Lemma B.1 (Equation (2) in Lemma 1 from Tibshirani et al. (2019))**.** *Let $v_1, \ldots, v_{n+1} \in \mathbb{R}$ and $(p_1, \ldots, p_{n+1}) \in \mathbb{R}$ be non-negative reals summing to 1. Then for any $\beta \in [0, 1]$ and*

$$v_{n+1} \le \text{Quantile}\left(\beta; \sum_{i=1}^{n+1} p_i\delta_{v_i}\right) \iff v_{n+1} \le \text{Quantile}\left(\beta; \sum_{i=1}^{n} p_i\delta_{v_i} + p_{n+1}\delta_\infty\right).$$

Lemma B.2 (Equation (10) from Tibshirani et al. (2019))**.** *Let $d_{\mathrm{TV}}(Q_{1X}, Q_{2X})$ denote the total-variation distance between Q_{1X} and Q_{2X}. Then*

$$d_{\mathrm{TV}}\left(Q_{1X} \times P_{Y|X}, Q_{2X} \times P_{Y|X}\right) = d_{\mathrm{TV}}\left(Q_{1X}, Q_{2X}\right)$$

Lemma B.3 (Theorem 2 of von Bahr et al. (1965))**.** *Let Z_i be independent mean-zero random variables. Then for any $\delta \in [0, 1)$,*

$$\mathbb{E}\left|\sum_{i=1}^n Z_i\right|^{1+\delta} \le 2\sum_{i=1}^n \mathbb{E}|Z_i|^{1+\delta}.$$

Throughout the proof, we let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$ be the calibration set and let $(X_{n+1}, Y_{n+1}) \sim Q_X \times P_{Y|X}$ be a test point. Further, let $Z_i = (X_i, Y_i)$, $Z = (Z_1, \ldots, Z_{n+1})$, and $V = (V_1, \ldots, V_{n+1})$. For any permutation π on $\{1, \ldots, n+1\}$ and $v^\star \in \mathbb{R}^{n+1}$, let $v^\star_\pi = (v^\star_{\pi(1)}, \ldots, v^\star_{\pi(n+1)})$. Additionally, we define $\mathcal{E}(v)$ to be the unordered set of v (where repetition of elements is allowed).

Proof of Theorem B.1 under assumption B1 We first consider the case where Q_X is absolutely continuous with respect to P_X, i.e.

$$\mathbb{P}_{X\sim Q_X}(w(X) < \infty) = 1.$$

In this case, for any measurable function f,

$$\mathbb{E}_{X\sim Q_X}[f(X)] = \mathbb{E}_{X\sim P_X}[w(X)f(X)]. \tag{B.3}$$

On the other hand, it always holds that $\mathbb{P}_{X\sim P_X}(w(X) < \infty) = 1$. In addition, the assumption $\mathbb{E}_{X\sim P_X}[\hat{w}(X) \mid \mathcal{Z}_{\mathrm{tr}}] < \infty$ implies that $\mathbb{P}_{X\sim P_X}(\hat{w}(X) < \infty) = 1$. By (B.3),

$$\mathbb{P}_{X\sim Q_X}(\hat{w}(X) < \infty) = 1 - \mathbb{E}_{X\sim P_X}[w(X)I(\hat{w}(X) = \infty)].$$

Since the integrand is non-negative,

$$\begin{aligned}&\mathbb{E}_{X\sim P_X}[w(X)I(\hat{w}(X) = \infty)] \\ &\quad = \lim_{K\to\infty} \mathbb{E}_{X\sim P_X}[w(X)I(w(X) \le K, \hat{w}(X) = \infty)] \le \lim_{K\to\infty} K\mathbb{P}_{X\sim P_X}(\hat{w}(X) = \infty) = 0.\end{aligned}$$

Thus, we also have

$$\mathbb{P}_{X\sim Q_X}(\hat{w}(X) < \infty) = 1.$$

Next, observe that

$$(V \mid \mathcal{E}(Z) = \mathcal{E}(z^\star), \mathcal{Z}_{\mathrm{tr}}) \overset{\mathrm{d}}{=} v^\star_\Pi,$$

where Π is a random permutation on $\{1, \ldots, n+1\}$ with

$$\begin{aligned}\mathbb{P}\left(\Pi = \pi \mid \mathcal{E}(Z) = \mathcal{E}(z^\star), \mathcal{Z}_{\mathrm{tr}}\right) =& \frac{\mathbb{P}(Z_1 = z^\star_{\pi(1)}, \ldots, Z_{n+1} = z^\star_{\pi(n+1)})}{\sum_{\pi'} \mathbb{P}(Z_1 = z^\star_{\pi'(1)}, \ldots, Z_{n+1} = z^\star_{\pi'(n+1)})} \\ =& \frac{w(x^\star_{\pi(n+1)})}{n! \sum_{i=1}^{n+1} w(x^\star_i)}.\end{aligned}$$

Note that this conditional probability is well-defined because $w(X) < \infty$ almost surely under both P_X and Q_X. Consequently, for any $i \in \{1, \ldots, n+1\}$,

$$\mathbb{P}\left(\Pi(n+1) = j \mid \mathcal{E}(Z) = \mathcal{E}(z^\star), \mathcal{Z}_{\mathrm{tr}}\right) = \frac{w(x^\star_j)}{\sum_{i=1}^{n+1} w(x^\star_i)} = p_j\left(x^\star_{n+1}\right),$$

where p_{n+1} refers to p_∞ for notational convenience. As a result,

$$(V_{n+1} \mid \mathcal{E}(Z) = \mathcal{E}(z^\star), \mathcal{Z}_{\mathrm{tr}}) \overset{\mathrm{d}}{=} \sum_{j=1}^{n+1} p_j(x^\star_{n+1}) \delta_{v^\star_j}.$$

Let $\widetilde{Q}_X$ be a new measure with

$$d\widetilde{Q}_X(x) = \hat{w}(x) dP_X(x).$$

Since $\mathbb{E}_{X\sim P_X}[\hat{w}(X)] = 1$, $\mathbb{P}_{X\sim P_X}(\hat{w}(X) < \infty) = 1$. As a result, $\tilde{Q}_X$ is a probability measure. Let $\widetilde{Z}_{n+1} = (\widetilde{X}_{n+1}, \widetilde{Y}_{n+1}) \sim \widetilde{Q}_X \times P_{Y|X}$. Further let $\widetilde{Z} = (Z_1, \ldots, Z_n, \widetilde{Z}_{n+1})$ and $\widetilde{V} = (\widetilde{V}_1, \ldots, \widetilde{V}_{n+1})$ denote the conformity score computed with $\widetilde{Z}$. Using the above result, we have that

$$\left(\widetilde{V}_{n+1} \mid \mathcal{E}(\widetilde{Z}), \mathcal{Z}_{\mathrm{tr}}\right) \overset{\mathrm{d}}{=} \sum_{j=1}^{n+1} \hat{p}_j(\widetilde{X}_{n+1}) \delta_{\widetilde{V}_j}.$$

Note that each $\hat{p}_i(\tilde{X}_{n+1})$, $i = 1, \ldots, n+1$ is well-defined since $\hat{w}(X_i)$ is almost surely finite under both P_X and Q_X. As a consequence,

$$
\begin{aligned}
&\mathbb{P}\left(\widetilde{Y}_{n+1} \in \hat{C}(\widetilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) \\
&= \mathbb{P}\left(\widetilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{p}_i(\widetilde{X}_{n+1})\delta_{\widetilde{V}_i} + \hat{p}_{n+1}(\widetilde{V}_{n+1})\delta_{\infty}\right) \mid \mathcal{Z}_{\mathrm{tr}}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\widetilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{p}_i(\widetilde{X}_{n+1})\delta_{\widetilde{V}_i} + \hat{p}_{n+1}(\widetilde{V}_{n+1})\delta_{\infty}\right) \mid \mathcal{E}(\widetilde{Z}), \mathcal{Z}_{\mathrm{tr}}\right) \mid \mathcal{Z}_{\mathrm{tr}}\right] \\
&\overset{(1)}{=} \mathbb{E}\left[\mathbb{P}\left(\widetilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n+1} \hat{p}_i(\widetilde{X}_{n+1})\delta_{\widetilde{V}_i}\right) \mid \mathcal{E}(\widetilde{Z}), \mathcal{Z}_{\mathrm{tr}}\right) \mid \mathcal{Z}_{\mathrm{tr}}\right] \\
&\overset{(2)}{\ge} 1-\alpha,
\end{aligned}
$$

where step (1) is due to Lemma B.1, and step (2) follows from the definition of the quantile function. On the other hand,

$$
\begin{aligned}
&\left|\mathbb{P}\left(\widetilde{Y}_{n+1} \in \hat{C}(\widetilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{ca}}, \mathcal{Z}_{\mathrm{tr}}\right) - \mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{Z}_{\mathrm{ca}}, \mathcal{Z}_{\mathrm{tr}}\right)\right| \\
&\qquad\qquad \le d_{\mathrm{TV}}\left(\widetilde{Q}_X \times P_{Y|X}, Q_X \times P_{Y|X}\right) = d_{\mathrm{TV}}\left(\widetilde{Q}_X, Q_X\right),
\end{aligned}
$$

where the last equality is a result of Lemma B.2. Rearranging the above inequality and taking expectation w.r.t. $\mathcal{Z}_{\mathrm{ca}}$, we obtain that

$$
\begin{aligned}
\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) &\ge \mathbb{P}\left(\widetilde{Y}_{n+1} \in \hat{C}(\widetilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) - d_{\mathrm{TV}}\left(\widetilde{Q}_X, Q_X\right) \\
&\ge 1-\alpha - d_{\mathrm{TV}}\left(\widetilde{Q}_X, Q_X\right).
\end{aligned}
$$

Note that

$$
d_{\mathrm{TV}}\left(\widetilde{Q}_X, Q_X\right) = \frac{1}{2}\int |d\widetilde{Q}_X - dQ_X| = \frac{1}{2}\int |\hat{w}(x) - w(x)| dP_X(x) = \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)|.
$$

With assumption B1, we reach the conclusion

$$
\lim_{N,n\to\infty} \mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}\left(Y \in \hat{C}(X)\right) = \lim_{N,n\to\infty} \mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1})\right) \ge 1-\alpha. \tag{B.4}
$$

Next, we extend the result to the case where $\mathbb{P}_{X\sim Q_X}(w(X) < \infty) < 1$. If $\mathbb{P}_{X\sim P_X}(\hat{w}(X) < \infty) < 1$, it is clear that $\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)| = \infty$, conflicting with the assumption B1. Thus, $\mathbb{P}_{X\sim P_X}(\hat{w}(X) < \infty) = 1$.

Let Q'_X denote the distribution Q_X conditional on the event $\mathcal{V}_\infty \triangleq \{x : w(x) < \infty\}$; that is,

$$
dQ'_X(x) = \frac{I(x \in \mathcal{V}_\infty) dQ_X(x)}{\mathbb{P}_{X\sim Q_X}(\mathcal{V}_\infty)}.
$$

Further, set $w'(x) = dQ'_X(x)/dP_X(x)$ and $\hat{w}'(x) = \hat{w}(x)I(w \in \mathcal{V}_\infty)/\mathbb{P}_{X\sim Q_X}(\mathcal{V}_\infty)$. Note that $\hat{C}(x)$ remains the same on $\mathcal{V}_\infty$ when $\hat{w}$ is replaced by $\hat{w}'$ and Q_X is replaced by Q'_X, because weighted split conformal inference is invariant with respect to rescalings of the covariate shift estimate. Since $\mathbb{P}_{X\sim Q'_X}(w(X) < \infty) = 1$, (B.4) implies that

$$
\mathbb{P}_{(X,Y)\sim Q'_X \times P_{Y|X}}\left(Y \in \hat{C}(X)\right) \ge 1-\alpha - \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}'(X) - w'(X)|.
$$

It can be reformulated as

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid w(X_{n+1}) < \infty\right) \geq 1 - \alpha - \frac{1}{2\mathbb{P}_{X\sim Q_X}(\mathcal{V}_\infty)}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)|.$$

On the other hand, when $w(X_{n+1}) = \infty$, $\eta(X_{n+1}) = \infty$ implying that $\hat{L}(X_{n+1}) = -\infty$. As a result,

$$\mathbb{P}\left(Y_{n+1} \geq \hat{L}(X_{n+1}) \mid w(X_{n+1}) = \infty\right) = 1.$$

Putting the two pieces together, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(Y_{n+1} \geq \hat{L}(X_{n+1})\right) \\
&= \mathbb{P}\left(Y_{n+1} \geq \hat{L}(X_{n+1}) \mid w(X_{n+1}) < \infty\right)\mathbb{P}\left(w(X_{n+1}) < \infty\right) \\
&\qquad + \mathbb{P}\left(Y_{n+1} \geq \hat{L}(X_{n+1}) \mid w(X_{n+1}) = \infty\right)\mathbb{P}\left(w(X_{n+1}) = \infty\right) \\
&\geq (1-\alpha)\mathbb{P}_{X\sim Q_X}(\mathcal{V}_\infty) + \mathbb{P}_{X\sim Q_X}(\mathcal{V}_\infty^c) - \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)| \\
&\geq 1 - \alpha - \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)| \\
&\to 1 - \alpha.
\end{aligned}$$

Proof of Theorem B.1 under Assumption B2 Note that Assumption B2 (b) (i) implies that $w(X)$ is almost surely finite under Q_X and $\hat{w}(X)$ is almost surely finite under P_X. By the same reasoning as in the last subsection, $w(X)$ is almost surely finite under P_X and $\hat{w}(X)$ is almost surely finite under Q_X.

Let $(\widetilde{X}, \widetilde{Y}) \sim Q_X \times P_{Y|X}$. Under Assumption B2, our first claim is that the calibration term $\eta(\widetilde{X})$ will be larger than s_0 asymptotically. This statement is formalized in Lemma B.4.

Lemma B.4. *Under Assumption B2, for any* $0 < \varepsilon < r/3$*,*

$$\lim_{N,n\to\infty}\mathbb{P}_{\widetilde{X}\sim Q_X}\left(\eta(\widetilde{X}) \geq s_0 - \varepsilon\right) = 1.$$

We defer the proof of Lemma B.4 to the end of this section. For now, assuming that Lemma B.4 holds, we have that for any $0 < \varepsilon < r/3$,

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{Y} \in \hat{C}(\widetilde{X}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&= \mathbb{P}\left(V(\widetilde{X}, \widetilde{Y}; \mathcal{Z}_{\mathrm{ca}}) \leq \eta(\widetilde{X}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\geq \mathbb{P}\left(\widetilde{Y} \in \mathcal{F}_{\eta(\widetilde{X})}(\widetilde{X}; \mathcal{Z}_{\mathrm{tr}}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\geq \mathbb{P}\left(\widetilde{Y} \in \mathcal{F}_{s_0-\varepsilon}(\widetilde{X}; \mathcal{Z}_{\mathrm{ca}}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) - \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\overset{(1)}{\geq} \mathbb{P}\left(\widetilde{Y} \in \mathcal{O}_{s_0-2\varepsilon-\Delta(\widetilde{X})}(\widetilde{X}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) - \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\geq \mathbb{P}\left(\widetilde{Y} \in \mathcal{O}_{s_0-2\varepsilon-\Delta(\widetilde{X})\mathbf{1}\{\Delta(\widetilde{X})\leq\varepsilon\}}(\widetilde{X}) \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) - \mathbf{1}\{\Delta(\widetilde{X}) > \varepsilon\} - \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\overset{(2)}{\geq} 1 - \alpha - b_2\left(2\varepsilon + \Delta(\widetilde{X})\mathbf{1}\{\Delta(\widetilde{X}) \leq \varepsilon\}\right) - \mathbf{1}\{\Delta(\widetilde{X}) > \varepsilon\} - \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right) \\
&\geq 1 - \alpha - 3b_2\varepsilon - \mathbf{1}\{\Delta(\widetilde{X}) > \varepsilon\} - \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon \mid \widetilde{X}, \mathcal{Z}_{\mathrm{tr}}\right).
\end{aligned}$$

Above, step (1) is due to the definition of $\Delta(x)$, and step (2) follows from condition *(b) (i)* in Assumption B2. Taking expectation w.r.t. $\widetilde{X}$ and $\mathcal{Z}_{\text{tr}}$,

$$
\begin{aligned}
\mathbb{P}\left(\widetilde{Y} \in \hat{C}(\widetilde{X})\right) &\geq 1-\alpha-3b_2\varepsilon-\mathbb{P}\left(\Delta(\widetilde{X})>\varepsilon\right)-\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\right) \\
&\overset{(1)}{\geq} 1-\alpha-3b_2\varepsilon-\mathbb{E}\left[\Delta(\widetilde{X})\right]/\varepsilon-\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\right) \\
&= 1-\alpha-3b_2\varepsilon-\mathbb{E}_{X\sim P_X}\left[\Delta(X)w(X)\right]/\varepsilon-\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\right),
\end{aligned}
$$

where step (1) is implied by Markov's inequality. By Assumption B2, the second term vanishes asymptotically. By Lemma B.4,

$$
\lim_{N,n\to\infty}\mathbb{P}_{(X,Y)\sim Q_X\times P_{Y\mid X}}\left(Y\in\hat{C}(X)\right)=\lim_{N,n\to\infty}\mathbb{P}\left(\widetilde{Y}\in\hat{C}(\widetilde{X})\right)\geq 1-\alpha-3b_2\varepsilon.
$$

Since the above holds for arbitrary $\varepsilon>0$, we complete the proof of (B.1).

To see (B.2), note that

$$
\begin{aligned}
&\mathbb{P}\left(\mathbb{P}\left(\widetilde{Y}\in\hat{C}(\widetilde{X})\mid\widetilde{X},\mathcal{Z}_{\text{tr}}\right)\leq 1-\alpha-4b_2\varepsilon\right) \\
&\quad\leq \mathbb{P}\left(1-\alpha-3b_2\varepsilon-\mathbf{1}\{\Delta(\widetilde{X})>\varepsilon\}-\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\mid\widetilde{X},\mathcal{Z}_{\text{tr}}\right)\leq 1-\alpha-4b_2\varepsilon\right) \\
&\quad\leq \mathbb{P}\left(\mathbf{1}\{\Delta(\widetilde{X})>\varepsilon\}+\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\mid\widetilde{X},\mathcal{Z}_{\text{tr}}\right)\geq b_2\varepsilon\right) \\
&\quad\leq \mathbb{P}\left(\Delta(\widetilde{X})>\varepsilon\right)+\mathbb{P}\left(\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\mid\widetilde{X},\mathcal{Z}_{\text{tr}}\right)>b_2\varepsilon\right) \\
&\quad\leq \mathbb{E}\left[\Delta(X)w(X)\right]/\varepsilon+\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\right)/(b_2\varepsilon),
\end{aligned}
$$

where the last step applies (B.3) and Markov's inequality. By condition *(b) (ii)* in assumption B2 and Lemma B.4,

$$
\lim_{N,n\to\infty}\mathbb{P}\left(\mathbb{P}\left(\widetilde{Y}\in\hat{C}(\widetilde{X})\mid\widetilde{X},\mathcal{Z}_{\text{tr}}\right)\leq 1-\alpha-4b_2\varepsilon\right)=0.
$$

Replacing $4b_2\varepsilon$ with ε yields the desired result.

Finally we present a proof of Lemma B.4.

Proof of Lemma B.4 Let $\mathcal{D}=\{\mathcal{Z}_{\text{tr}},(X_i)_{i=1}^n,\widetilde{X}\}$, $G(t)$ be the (random) CDF of $\sum_{i=1}^n\hat{p}_i(\widetilde{X})\delta_{V_i}+\hat{p}_\infty(\widetilde{X})\delta_{V_\infty}$ and $G^*(t)=\mathbb{E}[G(t)\mid\mathcal{D}]$. To begin with, note that

$$
\mathbb{P}\left(\eta(\widetilde{X})<s_0-\varepsilon\right)\leq\mathbb{P}\left(G(s_0-\varepsilon)\geq 1-\alpha\right).
$$

It then suffices to show that $\lim_{N,n\to\infty}\mathbb{P}(G(s_0-\varepsilon)\geq 1-\alpha)=0$ for all $0<\varepsilon\leq r/3$. By definition,

$$
G(s_0-\varepsilon)-G^*(s_0-\varepsilon)=\sum_{i=1}^n\hat{p}_i(\widetilde{X})\left[\mathbf{1}\{V_i\leq s_0-\varepsilon\}-\mathbb{P}(V_i\leq s_0-\varepsilon\mid\mathcal{D})\right].
$$

Conditional on $\mathcal{D}$, $G(s_0-\varepsilon)-G^\star(s_0-\varepsilon)$ is σ^2-sub-Gaussian, where

$$
\sigma^2=\sum_{i=1}^n\hat{p}_i^2(\widetilde{X})\leq\max_{1\leq i\leq n}\hat{p}_i(\widetilde{X})\left(\sum_{i=1}^n\hat{p}_i(\widetilde{X})\right)\leq\max_{1\leq i\leq n}\hat{p}_i(\widetilde{X}).
$$

Then by Hoeffding's inequality,

$$\mathbb{P}\left(G(s_0-\varepsilon)-G^*(s_0-\varepsilon)\geq t \mid \mathcal{D}\right)\leq \exp\left(-\frac{t^2}{2\max_i \hat{p}_i(\widetilde{X})}\right) \tag{B.5}$$

On the other hand, when $\varepsilon < r/3$,

$$\begin{aligned}
G^*(s_0-\varepsilon) &= \sum_{i=1}^n \hat{p}_i(\widetilde{X})\mathbb{P}\left(V_i \leq s_0-\varepsilon \mid \mathcal{D}\right)\\
&= \sum_{i=1}^n \hat{p}_i(\widetilde{X})\left[\mathbb{P}\left(V_i \leq s_0-\varepsilon, \Delta(X_i)>\varepsilon/2 \mid \mathcal{D}\right)+\mathbb{P}\left(V_i \leq s_0-\varepsilon, \Delta(X_i)\leq\varepsilon/2 \mid \mathcal{D}\right)\right]\\
&\overset{(1)}{\leq} \sum_{i=1}^n \hat{p}_i(\widetilde{X})\left[\mathbf{1}\left\{\Delta(X_i)>\varepsilon/2\right\}+\mathbb{P}\left(V_i \leq s_0-\varepsilon/2-\Delta(X_i), \Delta(X_i)\leq\varepsilon/2 \mid \mathcal{D}\right)\right]\\
&\leq \sum_{i=1}^n \hat{p}_i(\widetilde{X})\left[\mathbf{1}\left\{\Delta(X_i)>\varepsilon/2\right\}+\mathbb{P}\left(Y_i \in \mathcal{F}_{s_0-\varepsilon/4-\Delta(X_i)}(X_i;\mathcal{Z}_{\mathrm{tr}}) \mid \mathcal{D}\right)\right]\\
&\overset{(2)}{\leq} \sum_{i=1}^n \hat{p}_i(\widetilde{X})\left[\mathbf{1}\left\{\Delta(X_i)>\varepsilon/2\right\}+\mathbb{P}\left(Y_i \in \mathcal{O}_{s_0-\varepsilon/8}(X_i) \mid \mathcal{D}\right)\right]\\
&\overset{(3)}{\leq} \sum_{i=1}^n \hat{p}_i(\widetilde{X})\mathbf{1}\left\{\Delta(X_i)>\varepsilon/2\right\}+1-\alpha-\frac{b_1\varepsilon}{8}\\
&\leq \frac{2}{\varepsilon}\sum_{i=1}^n \hat{p}_i(\widetilde{X})\Delta(X_i)+1-\alpha-\frac{b_1\varepsilon}{8},
\end{aligned} \tag{B.6}$$

where step (1) holds because $\Delta(X_i)$ is deterministic conditional on $\mathcal{D}$, step (2) follows from the definition of $\Delta(X_i)$, and step (3) follows from the condition *(b) (i)*. Combining (B.5) and (B.6), we have

$$\begin{aligned}
&\mathbb{P}(G(s_0-\varepsilon)\geq 1-\alpha)\\
&\quad\leq \mathbb{P}\left(G(s_0-\varepsilon)-G^*(s_0-\varepsilon)\geq\frac{b_1\varepsilon}{16}\right)+\mathbb{P}\left(\sum_{i=1}^n \hat{p}_i(\widetilde{X})\Delta(X_i)\geq\frac{b_1\varepsilon^2}{32}\right)\\
&\quad\leq \mathbb{E}\left[\exp\left(-\frac{b_1^2\varepsilon^2}{512\max_i \hat{p}_i(\widetilde{X})}\right)\right]+\mathbb{P}\left(\sum_{i=1}^n \hat{p}_i(\widetilde{X})\Delta(X_i)\geq\frac{b_1\varepsilon^2}{32}\right)\\
&\quad\leq \exp\left(-\frac{b_1^2\varepsilon^2}{512}\log n\right)+\mathbb{P}\left(\max_i \hat{p}_i(\widetilde{X})\geq\frac{1}{\log n}\right)+\mathbb{P}\left(\sum_{i=1}^n \hat{p}_i(\widetilde{X})\Delta(X_i)\geq\frac{b_1\varepsilon^2}{32}\right)
\end{aligned} \tag{B.7}$$

By definition, for any value of $\widetilde{X}$,

$$\hat{p}_i(\widetilde{X})\leq\frac{\hat{w}(X_i)}{\sum_{j=1}^n \hat{w}(X_j)}.$$

Consequently,

$$\mathbb{P}\left(\max_i \hat{p}_i(\widetilde{X})\geq\frac{1}{\log n}\right)\leq\mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i)\leq\frac{n}{2}\right)+\mathbb{P}\left(\max_i \hat{w}(X_i)\geq\frac{n}{2\log n}\right).$$

Assume without loss of generality that $0 < \delta < 1$. Since $\mathbb{E}[\hat{w}(X_i) \mid \mathcal{Z}_{\text{tr}}] = 1$, Markov's inequality and Lemma B.3 give

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i) \le \frac{n}{2}\right) \le \mathbb{P}\left(\left|\sum_{i=1}^n \hat{w}(X_i) - 1\right| \ge \frac{n}{2}\right) \le \frac{2^{1+\delta}}{n^{1+\delta}}\mathbb{E}\left[\left|\sum_{i=1}^n \hat{w}(X_i) - 1\right|^{1+\delta}\right] \\
&\le \frac{2^{2+\delta}}{n^{1+\delta}}\sum_{i=1}^n \mathbb{E}\left[|\hat{w}(X_i) - 1|^{1+\delta}\right] = \frac{2^{2+\delta}}{n^{\delta}}\mathbb{E}\left[|\hat{w}(X_1) - 1|^{1+\delta}\right] \le \frac{2^{2+2\delta}}{n^{\delta}}\left(\mathbb{E}\left[\hat{w}(X_1)^{1+\delta}\right] + 1\right). \quad \text{(B.8)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}\left(\max_i \hat{w}(X_i) \ge \frac{n}{2\log n}\right) \le \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{1+\delta}}\mathbb{E}\left[\left(\max_i \hat{w}(X_i)\right)^{1+\delta}\right] \quad \text{(B.9)} \\
&\le \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{1+\delta}}\mathbb{E}\left[\sum_{i=1}^n \hat{w}(X_i)^{1+\delta}\right] = \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{\delta}}\mathbb{E}[\hat{w}(X_i)^{1+\delta}].
\end{aligned}$$

Combining (B.8) and (B.9) yields

$$\lim_{N,n\to\infty} \mathbb{P}\left(\max_i \hat{p}_i(\widetilde{X}) \ge \frac{1}{\log n}\right) = 0. \quad \text{(B.10)}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n \hat{p}_i(\widetilde{X})\Delta(X_i) \ge \frac{b_1\varepsilon^2}{32}\right) \le \mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i) \le \frac{n}{2}\right) + \mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i)\Delta(X_i) \ge \frac{b_1 n\varepsilon^2}{64}\right) \\
\le &\mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i) \le \frac{n}{2}\right) + \frac{64}{b_1\varepsilon}\mathbb{E}\left[\hat{w}(X_i)\Delta(X_i)\right] \overset{N,n\to\infty}{\longrightarrow} 0. \quad \text{(B.11)}
\end{aligned}$$

Together, (B.7), (B.10) and (B.11) imply

$$\lim_{N,n\to\infty} \mathbb{P}\left(\eta(\widetilde{X}) < s_0 - \varepsilon\right) = 0.$$

B.3 Proof of Theorem 2

Recall that $Y_i = T_i \wedge c_0$, $w(x) = 1/c(x)$ and $\hat{w}(x) = 1/\hat{c}(x)$. First we show that Assumption A1 of Theorem 2 implies Assumption B1 of Theorem B.1. Since $\mathbb{E}[1/\hat{c}(X) \mid \mathcal{Z}_{\text{tr}}] < \infty$ almost surely and $\mathbb{E}[1/c(X)] < \infty$, we set

$$\hat{w}_N(x) = \frac{1/\hat{c}(x)}{\mathbb{E}[1/\hat{c}(X) \mid \mathcal{Z}_{\text{tr}}]}, \ w(x) = \frac{dP_X(x)}{dP_{X|T=1}(x)} = \frac{1/c(x)}{\mathbb{E}[1/c(X)]}$$

and observe

$$\mathbb{E}[\hat{w}_N(X) \mid \mathcal{Z}_{\text{tr}}] = 1 = \mathbb{E}[w(X)].$$

Thus, Assumption B1 reduces to

$$\lim_{N\to\infty} \mathbb{E}\left|\frac{1/\hat{c}(X)}{\mathbb{E}[1/\hat{c}(X) \mid \mathcal{Z}_{\text{tr}}]} - \frac{1/c(X)}{\mathbb{E}[1/c(X)]}\right| = 0.$$

In fact,

$$
\begin{aligned}
&\lim_{N\to\infty}\mathbb{E}\left|\frac{1/\hat{c}(X)}{\mathbb{E}[1/\hat{c}(X)\mid \mathcal{Z}_{\mathrm{tr}}]}-\frac{1/c(X)}{\mathbb{E}[1/c(X)]}\right|\\
&\le \limsup_{N\to\infty}\frac{1}{\mathbb{E}[1/\hat{c}(X)\mid \mathcal{Z}_{\mathrm{tr}}]}\mathbb{E}\left|\frac{1}{\hat{c}(X)}-\frac{1}{c(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{c(X)}\right]\mathbb{E}\left|\frac{1}{\mathbb{E}[1/\hat{c}(X)\mid \mathcal{Z}_{\mathrm{tr}}]}-\frac{1}{\mathbb{E}[1/c(X)]}\right|\\
&\overset{(1)}{\le} \limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{c}(X)}-\frac{1}{c(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{c(X)}\right]\mathbb{E}\left|\mathbb{E}\left[\frac{1}{\hat{c}(X)}\mid \mathcal{Z}_{\mathrm{tr}}\right]-\mathbb{E}\left[\frac{1}{c(X)}\right]\right|\\
&\le \limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{c}(X)}-\frac{1}{c(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{c(X)}\right]\mathbb{E}\left(\mathbb{E}\left[\left|\frac{1}{\hat{c}(X)}-\frac{1}{c(X)}\right|\mid \mathcal{Z}_{\mathrm{tr}}\right]\right)\\
&\overset{(2)}{=} \left(1+\mathbb{E}\left[\frac{1}{c(X)}\right]\right)\limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{c}(X)}-\frac{1}{c(X)}\right|=0,
\end{aligned}
$$

where step (1) uses the fact that $c(x), \hat{c}(x) \in [0,1]$, and step (2) uses Assumption A1 and the condition $\mathbb{E}[1/c(X)] < \infty$.

Next we show that Assumption A2 implies Assumption B2. Remark that Assumption B2 (a) is satisfied since $\mathbb{E}[1/\hat{c}(X)^{1+\delta}] < \infty$). It remains to prove that A2 $\Longrightarrow$ B2 (b). Let

$$\mathcal{O}_s(x) = [q_\alpha(x; c_0) - s, \infty).$$

Clearly, Assumption A2 (i) implies Assumption B2 (b) (i) with $s_0 = 0$. Moreover,

$$
\begin{aligned}
\Delta_s(x) &= \inf\left\{\Delta : \hat{q}_\alpha(x;c_0) - s + \Delta \ge q_\alpha(x;c_0) - s \text{ and } q_\alpha(x;c_0) - s + \Delta \ge \hat{q}_\alpha(x;c_0) - s\right\}\\
&= \inf\left\{\Delta : \hat{q}_\alpha(x;c_0) + \Delta \ge q_\alpha(x;c_0) \text{ and } q_\alpha(x;c_0) + \Delta \ge \hat{q}_\alpha(x;c_0)\right\}\\
&= |\hat{q}_\alpha(x;c_0) - q_\alpha(x;c_0)| = \mathcal{E}(x).
\end{aligned}
$$

Thus, $\Delta(x) = \sup_{s\in[s_0-r,s_0]} \Delta_s(x) = \mathcal{E}(x)$ and Assumption B2 (b) (ii) holds.

B.4 Proof of Theorem 3

Using the same argument as in the proof of Theorem 2, it suffices to show that A2 from Theorem 3 implies B2 (b) from Theorem B.1. Let

$$\mathcal{O}_s(x) = [q_{\alpha-s}(x; c_0), \infty),$$

Clearly, Assumption A2 (i) implies Assumption B2 (b) (i) with $s_0 = 0$.

To compute $\Delta(x)$, we first replace r by $2r$. Assumption A2 (i) with $\varepsilon = 0$ implies that $F(q_{\alpha-s}(x;c_0) \mid X = x) = 1 - \alpha + s$ for any $s \in [s_0 - 2r, s_0] = [-2r, 0]$. Then for any $s \in [-r, 0]$,

$$\Delta_s(x) = \inf\left\{\Delta : q_{\alpha-s+\Delta}(x;c_0) \ge \hat{q}_{\alpha-s}(x;c_0) \text{ and } \hat{q}_{\alpha-s+\Delta}(x;c_0) \ge q_{\alpha-s}(x;c_0)\right\}.$$

Let $\mathcal{V}$ denote the event that $\mathcal{E}(X) \le r$. Then on $\mathcal{V}$, $\alpha - s + \mathcal{E}(X) \in [\alpha, \alpha + 2r]$, and by A2 (i),

$$
\begin{aligned}
&\alpha - s + \mathcal{E}(X) = F(q_{\alpha-s}(X;c_0) \mid X) + \mathcal{E}(X) \text{ and } F(q_{\alpha-s+\mathcal{E}(x)}(X;c_0) \mid X) = \alpha - s + \mathcal{E}(X)\\
&\Longrightarrow \alpha - s + \mathcal{E}(X) \ge F(\hat{q}_{\alpha-s}(X;c_0) \mid X) \text{ and } F(\hat{q}_{\alpha-s+\mathcal{E}(X)}(X;c_0) \mid X) \ge \alpha - s\\
&\Longrightarrow q_{\alpha-s+\mathcal{E}(X)}(X;c_0) \ge \hat{q}_{\alpha-s}(X;c_0) \text{ and } \hat{q}_{\alpha-s+\mathcal{E}(X)}(X;c_0) \ge q_{\alpha-s}(X;c_0)\\
&\Longrightarrow \Delta_s(X) \le \mathcal{E}(X).
\end{aligned}
$$

Thus on $\mathcal{V}$,

$$\Delta(X) = \sup_{s\in[s_0-r,s_0]} \Delta_s(X) \le \mathcal{E}(X).$$

On the other hand, $\Delta(X) \le 1$ almost surely. Since $c(X) \le 1$ almost surely, A2 (ii) implies that $\mathbb{E}[\Delta(X)] \to 0$. By Markov's inequality,

$$\mathbb{P}(\mathcal{V}) \to 1.$$

As a result,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[\Delta(X)/\hat{c}(X)] &\le \limsup_{N\to\infty} \mathbb{E}[\Delta(X) I_{\mathcal{V}}/\hat{c}(X)] + \mathbb{E}[\Delta(X) I_{\mathcal{V}^c}/\hat{c}(X)] \\
&\le \limsup_{N\to\infty} \mathbb{E}[\mathcal{E}(X)/\hat{c}(X)] + \alpha \mathbb{E}[I_{\mathcal{V}^c}/\hat{c}(X)] \\
&\le \limsup_{N\to\infty} \mathbb{E}[\mathcal{E}(X)/\hat{c}(X)] + \alpha \mathbb{P}(\mathcal{V}^c)^{1+1/\delta} \mathbb{E}[1/\hat{c}(X)^{1+\delta}] = 0,
\end{aligned}$$

where the second last step follows from Hölder's inequality. Similarly, we have

$$\lim_{N\to\infty} \mathbb{E}[\Delta(X)/c(X)] = 0.$$

Since $\mathbb{E}[1/\hat{c}(X)], \mathbb{E}[1/c(X)] \ge 1$, we conclude that

$$\lim_{N\to\infty} \mathbb{E}[\hat{w}(X)\Delta(X)] = \lim_{N\to\infty} \mathbb{E}[w(X)\Delta(X)] = 0.$$

In conclusion, A2 implies B2 (b) (ii). This completes the proof of Theorem 3.

B.5 When is CMR-LPB doubly robust?

For CMR, a natural oracle nested set is given by

$$\mathcal{O}_s(x) = [m(x; c_0) - s, \infty), \quad \text{where } m(x; c_0) = \mathbb{E}[T \wedge c_0 \mid X = x].$$

However, Assumption B2 (b) (i) with $\varepsilon = 0$ requires the existence of s_0 such that

$$\mathbb{P}(T \wedge c_0 \in \mathcal{O}_{s_0}(X) \mid X) = 1 - \alpha, \quad \text{almost surely.}$$

This implies that for some s_0,

$$\mathbb{P}(T \wedge c_0 \ge m(X; c_0) - s_0 \mid X) = 1 - \alpha, \quad \text{almost surely.}$$

The above equality does not hold in general. One exception is the additive case with homoscedastic errors:

$$T \wedge c_0 = m(X; c_0) + \nu, \quad \mathrm{Var}[\nu \mid X] = \mathrm{Var}[\nu].$$

In this case, we can derive the double robustness of the CMR-LPB based on Theorem B.1.