Journal of the Royal Statistical Society Series B:
Statistical Methodology, 2023, **85**, 24–45
https://doi.org/10.1093/jrsssb/qkac004
Advance access publication 28 January 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Conformalized survival analysis

Emmanuel Candès[1,2], Lihua Lei[2,3], and Zhimei Ren[4]

[1]Department of Mathematics, Stanford University, Stanford, CA, USA
[2]Department of Statistics, Stanford University, Stanford, CA, USA
[3]Graduate School of Business, Stanford University, Stanford, CA, USA
[4]Department of Statistics, University of Chicago, Chicago, IL, USA

Address for correspondence: Zhimei Ren, Jones 211, 5747 South Ellis Avenue, Chicago, IL 60637, USA. Email: zmren@uchicago.edu

Abstract

In this paper, we develop an inferential method based on 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. The validity and efficiency of our procedure are demonstrated on synthetic data and real COVID-19 data from the UK Biobank.

Keywords: censoring, distribution boosting, prediction interval, random forests, survival time, weighted conformal inference

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 individual risk assessment for patients who test positive; see Table 2 of Emanuel et al. (2020). Clearly, one risk 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 & 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 & Prentice, 2011). More often than not, the scientist has available multiple categorical and continuous

Received: April 15, 2021. **Revised:** April 18, 2022. **Accepted:** September 27, 2022

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 (Breslow, 1975; Cox, 1972). Other popular models, such as accelerated failure time (AFT) (Cox, 1972; Wei, 1992) and proportional odds models (Harrell, 2015; Murphy et al., 1997), also combine non-parametric 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., Faraggi & Simon, 1995; Goeman, 2010; Gui & Li, 2005; Hothorn et al., 2006; Ishwaran et al., 2008; Katzman et al., 2016; Lao et al., 2017; Li & Bradic, 2020; Simon et al., 2011; R. Tibshirani, 1997; Verweij & Van Houwelingen, 1993; Wang et al., 2019; Witten & Tibshirani, 2010; Zhang & Lu, 2007). An appealing feature of these methods is that they typically do not make modelling assumptions. To quote from Efron (2020): 'Neither surface nor noise is required as input to randomForest, gbm, or their kin'. 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 & Tibshirani, 1994), can yield unreliable results since their validity (1) rests on implicit modelling assumptions, and (2) holds only asymptotically (e.g., L. Lei & Candès, 2021; Ratkovic & 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., Aitchison & Dunsmore, 1980; Geisser, 1993; Krishnamoorthy & Mathew, 2009; Stine, 1985; Vovk et al., 2005; Wald, 1943; Wilks, 1941) 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., J. Lei & Wasserman, 2014; Saunders et al., 1999; R. J. Tibshirani et al., 2019; Vovk, 2002; Vovk et al., 2005). While coverage is only guaranteed in a marginal sense, it has been theoretically proved and 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., L. Lei & Candès, 2021; Sesia & 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 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 prediction 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 (R. J. 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 (Hong & Tamer, 2003; Powell, 1986; Sant'Anna, 2016).

2 Prediction intervals for survival times

2.1 Problem set-up

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 ith 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 ith unit include X_i, C_i, and the censored survival time $\tilde{T}_i$, defined as the minimum of the survival and censoring time:

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

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 $\tilde{T}_i = T_i$ if the ith patient died before the day of data analysis and $\tilde{T}_i = C_i$ 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 & Prentice, 2011):

Assumption 1 (conditionally independent censoring)

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

This assumes away any unmeasured confounder affecting both the survival and censoring time; 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:

Assumption 2 (completely independent censoring)

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

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., Anisimov & Fedorov, 2007; Barnard et al., 2010; Carter et al., 2005), and thus (1) holds as soon as the vector of covariates includes the site of the trial. For an observational study such as the COVID-19 example discussed later in Section 5, additional covariates would be included to make the conditionally independent censoring assumption plausible.

Although (1) is a strong assumption, it is a widely used starting point to study survival analysis methods (Kalbfleisch & Prentice, 2011). We leave the investigation of informative censoring (e.g., Lagakos, 1979; Scharfstein & Robins, 2002; Wu & Carroll, 1988) to future research. Additionally, whereas the setting of Type I censoring appears to be restrictive, we will show in Section 6.1 that an LPB in this setting can still be informative for other censoring types.

2.2 Naive lower prediction 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 $(X_i, C_i, \tilde{T}_i)_{i=1}^n$. We say an LPB is *calibrated* if it satisfies the following coverage criterion:

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

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 $\tilde{T} \leq T$, any calibrated LPB on the censored survival time $\tilde{T}$ 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 $\tilde{T}$ directly. Since the samples $(X_i, \tilde{T}_i)$ are i.i.d., a distribution-free calibrated LPB on $\tilde{T}$ can be obtained via standard techniques from conformal inference (e.g., J. Lei et al., 2018; Romano, Patterson, et al., 2019; Vovk et al., 2005). Our first result is somewhat negative: indeed, it states that all distribution-free calibrated LPBs on T must be LPBs on $\tilde{T}$.

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. Then for any such distribution,

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

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

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

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 $\tilde{T}$ 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 $\tilde{T}$, 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,

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

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

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

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

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

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

where (a) uses the fact that

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

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

$$\frac{\mathrm{d}P_X}{\mathrm{d}P_{X|C\geq c_0}}(x) = \frac{\mathbb{P}(C \geq c_0)}{\mathbb{P}(C \geq c_0 \mid X = x)}. \tag{6}$$

While there is a distributional shift between the selected units and the target population, the special form of the covariate shift allows us to adjust for the bias by carefully reweighting the samples. In particular, applying the one-sided version of weighted conformal inference (R. J. 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., J. Lei et al., 2018; Romano, Patterson, et al., 2019; Vovk et al., 2005). 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.

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 Sections 3.1 and 3.2, and discuss a data-adaptive approach to choosing this threshold in Section 3.4.

To deal with covariate shifts, R. J. Tibshirani et al. (2019) introduced weighted conformal inference, which extends standard conformal inference (e.g., Barber et al., 2019a, 2019b; Cauchois et al., 2020; J. Lei & Wasserman, 2014; Romano et al., 2020; Sadinle et al., 2019; Shafer & Vovk, 2008; Vovk et al., 2005). 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 prediction intervals for test points drawn from the target distribution $Q_X \times P_{Y|X}$ (in standard conformal inference, $P_X = Q_X$). Assuming $w(x) = \mathrm{d}Q_X(x)/\mathrm{d}P_X(x)$ is known, then weighted conformal inference produces prediction intervals $\hat{C}(\cdot)$ with the property

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

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).

In Algorithm 1, we sketched a version of weighted conformal inference based on data splitting, which is adapted to our setting and has low computational overhead. Operationally, it has three main steps:

(a) split the data into a training and a calibration fold;
(b) 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; here, we generate a conformity score such that a large value indicates a lack of conformity to training data.
(c) 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

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

Algorithm 1: conformalized survival analysis

```
Input: level $\alpha$; data $\mathcal{Z} = (X_i, \tilde{T}_i, C_i)_{i \in \mathcal{I}}$; 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;
function $\mathcal{C}(\mathcal{D})$ to select the threshold $c_0$ 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. Select $c_0 = \mathcal{C}(\mathcal{Z}_{\text{tr}})$ and let $\mathcal{I}'_{\text{ca}} = \{i \in \mathcal{I}_{\text{ca}} : C_i \geq c_0\}$.
3. For each $i \in \mathcal{I}'_{\text{ca}}$, compute the conformity score $V_i = V(X_i, \tilde{T}_i \wedge c_0; \mathcal{Z}_{\text{tr}})$.
4. For each $i \in \mathcal{I}'_{\text{ca}}$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{\text{tr}}) \in [0, \infty)$.
5. 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 $\hat{p}_\infty(x) = \frac{\hat{w}(x; \mathcal{Z}_{\text{tr}})}{\sum_{i \in \mathcal{I}'_{\text{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\text{tr}})}$.
6. Compute $\eta(x) = \mathsf{Quantile}(1 - \alpha; \sum_{i \in \mathcal{I}'_{\text{ca}}} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty)$.

Output: $\hat{L}(x) = \inf\{y : V(x, y; \mathcal{Z}_{\text{tr}}) \leq \eta(x)\} \wedge c_0$
```

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 $w(x) = 1/\hat{c}(x)$ in our case where $\hat{c}(x)$ is an estimate of $c(x)$. Finally, apart from fitting $V(\cdot, \cdot; \mathcal{Z}_{\text{tr}})$ and $\hat{w}(\cdot; \mathcal{Z}_{\text{tr}})$ once on the training fold, the additional computational cost of our algorithm comes from computing $|\mathcal{I}'_{\text{ca}}|$ conformity scores and finding the $(1 - \alpha)$th quantile. We provide a detailed analysis of time complexity in Section D.4 of the Supplementary material.

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 $(\hat{m}(x) - \eta(x)) \wedge c_0$. This is the one-sided version of the conformity score used in Vovk et al. (2005) and J. 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 αth quantile of Y given X. The resulting LPB is then $(\hat{q}_\alpha(x) - \eta(x)) \wedge c_0$. This score was proposed by Romano, Patterson, et al. (2019); 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 $\hat{F}^{-1}_{Y|X=x}(\alpha - \eta(x)) \wedge c_0$, 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, $\mathbb{P}(C \geq c_0 \mid X) = \mathbb{P}(C \geq c_0)$ 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 R. J. Tibshirani et al., 2019)

Let c_0 be any threshold independent of $\mathcal{Z}_{\text{ca}}$. Consider Algorithm 3.1 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 prediction 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 (J. H. Friedman, 2020) to estimate $c(x) = \mathbb{P}(C \geq c_0 \mid X = x)$. For two-sided weighted split-CQR, L. Lei and Candès (2021) 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 Section B in the Supplementary material, we present more general results, both non-asymptotic and asymptotic, that are applicable 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 of the asymptotic result tailored to the CQR-LPB for simplicity.

Theorem 2. Let $N = |\mathcal{Z}_{\text{tr}}|$, $n = |\mathcal{Z}_{\text{ca}}|$, c_0 be any threshold independent of $\mathcal{Z}_{\text{ca}}$, and $q_\alpha(x; c_0)$ denote the αth conditional quantile of $T \wedge c_0$ given $X = x$. Further, let $\hat{c}(x)$ and $\hat{q}_\alpha(x; c_0)$ be estimates of $c(x)$ and $q_\alpha(x; c_0)$ respectively using $\mathcal{Z}_{\text{tr}}$, and $\hat{L}(x)$ be the corresponding CQR-LPB. Assume that there exists $\delta > 0$ such that $\mathbb{E}[1/\hat{c}(X)^{1+\delta}] < \infty$ 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 $\varepsilon \in [0, r]$,

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

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

Then

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

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

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

Remark 1 The condition A2(i) holds if T has a bounded and absolutely continuous density conditional on X in a neighbourhood of $q_\alpha(x)$. In fact, noting that $q_\alpha(x; c_0) = q_\alpha(x) \wedge c_0$, when $q_\alpha(x; c_0) + \varepsilon \leq c_0$, we have $q_\alpha(x) \leq c_0$ and thus $T \wedge c_0 \geq q_\alpha(x) \wedge c_0$ if and only if $T \geq q_\alpha(x)$.

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

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

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 A2.

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 A2 is replaced by the following conditions:

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

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

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

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

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 prediction intervals derived from the Cox model (Kalbfleisch & Prentice, 2011); if the model is misspecified, however, the LPB is still calibrated.

Remark 2 A special case is when the completely independent censoring assumption holds, yet the researcher is unaware of this and still applies the estimated $\hat{c}(\cdot)$ to obtain the prediction intervals. As implied by Theorems 2 and 3, if $\hat{c}(\cdot)$ is approximately a constant function, the prediction interval is approximately calibrated. Notably, even if $\hat{c}(\cdot)$ deviates from a constant, our prediction interval still achieves coverage as long as the estimated weights are non-decreasing in the conformity scores. We present this additional robustness result in Section D.3 of the Supplementary material.

As a concluding remark, the prediction interval can become numerically and statistically unstable in the presence of extreme weights since the proposed method depends on $c(x)$ (or the estimated $\hat{c}(x)$) through its inverse. The reader may have observed that $c(x)$ plays a role similar to that of the propensity score in causal inference; the reweighting step in Algorithm 1 is analogous to inverse propensity score weighting-type methods. Assumption A1 in Theorem 2 mimics the overlap condition (e.g., D'Amour et al., 2021) in the causal inference literature. That said, there is a crucial difference. In a typical causal setting, the overlap condition is an assumption about the unknown data generating process, which cannot be manipulated. In contrast, in our work Assumption A1 can always be satisfied by selecting a sufficiently low threshold c_0. We provide a detailed discussion in Section D.1 of the Supplementary material.

3.3 Adaptivity to high-quality modelling

We have seen that when the quantiles of survival times are well estimated, $\hat{L}(x) \approx q_\alpha(x;\, c_0)$, which is the oracle lower prediction bound for $T \wedge c_0$, had the true survival function been known. This holds without knowing whether the survival model is well estimated or not. This suggests that conformalized survival analysis has favourable adaptivity properties, as formalized below.

Theorem 4. (a) Under the settings and assumptions of Theorem 2, assume further that A2(ii) holds and a modified version of A2 (i) holds: there exists $b_1 > 0$ and $r > 0$ such that, for any x and $\varepsilon \in [0, r]$,

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

Then, for any $\varepsilon > 0$,

$$\lim_{N,n\to\infty} \mathbb{P}_{X\sim Q_X}(\hat{L}(X) \geq q_\alpha(X; c_0) - \varepsilon) = 1.$$

(1) (b)] Under the settings and assumptions of Theorem 3, assume further the condition (ii) and the modified version of condition (i): there exists $r > 0$ such that, for any x and $\varepsilon \in [0, r]$,

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

Then, for any $\varepsilon > 0$,

$$\lim_{N,n\to\infty} \mathbb{P}_{X\sim Q_X}(\hat{L}(X) \geq q_{\alpha-\varepsilon}(X; c_0)) = 1.$$

In theory, if c_0 is allowed to grow with n and C exceeds c_0 with sufficient probability, then $\hat{L}(x) \approx q_\alpha(x)$ (see Supplementary material, Appendix C.3). In practice, it would however be wiser to tune c_0 in a data-adaptive fashion (discussed in the next subsection) than to prescribe a predetermined growing sequence.

3.4 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, Theorems 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.

Under suitable conditions, we can choose c_0 by using the calibration fold $\mathcal{Z}_{\text{ca}}$ and have the resulting LPBs still be (approximately) calibrated. To be specific, given a candidate set $\mathcal{C}$ for c_0, we simply maximize the average LPB on $\mathcal{Z}_{\text{ca}}$:

$$\hat{c}_0 = \operatorname*{argmax}_{c_0 \in \mathcal{C}} \frac{1}{|\mathcal{I}_{\text{ca}}|} \sum_{i \in \mathcal{I}_{\text{ca}}} \hat{L}_{c_0}(X_i),$$

where $\hat{L}_{c_0}(X)$ is given by the conformalized survival analysis with the threshold c_0. In Section D.2 of the Supplementary material we derive uniform results for the c_0's in $\mathcal{C}$, and prove coverage guarantees for $L_{\hat{c}_0}(X)$ via a generalization of the techniques for unweighted conformal inference by Yang and Kuchibhotla (2021).

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 αth 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 αth 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 & 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, Patterson, et al., 2019) naively to $(X_i, \tilde{T}_i)_{i=1}^n$, 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 (J. H. Friedman, 2020); for the CDR-LPB, the conditional survival function is estimated via distribution boosting, which is implemented in the R package `conTree` (J. Friedman & 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 training set is used for fitting the predictive model, and the remaining 50% of the training set is reserved for calibration. The splitting ratio between the training set and the test set is slightly different from the recommendation by 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. 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.

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

Table 1. Parameters used in the simulation study

	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	$\mathcal{U}([-1, 1]^p)$	$\mathcal{E}(0.4)$	$\log 2 + 1 + 0.55(x_1^2 - x_3x_5)$	$\lvert x_{10}\rvert + 1$

Note. '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.

We consider 2×2 settings with univariate or multivariate covariates plus homoscedastic or heteroscedastic errors. Here the term 'homoscedastic' or 'heteroscedastic' is applied to $\log T$. 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.

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 2a, 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 $\mathrm{Var}(T \mid X)$ 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, $\mathrm{Var}(T \mid X)$ 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 multivariate case, the conditional coverage is slightly uneven, though still concentrating around the target line. Figure 2b presents the ratio between the LPBs and the true αth conditional quantile as a function of Var(T | 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 21 January 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 21 January 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.

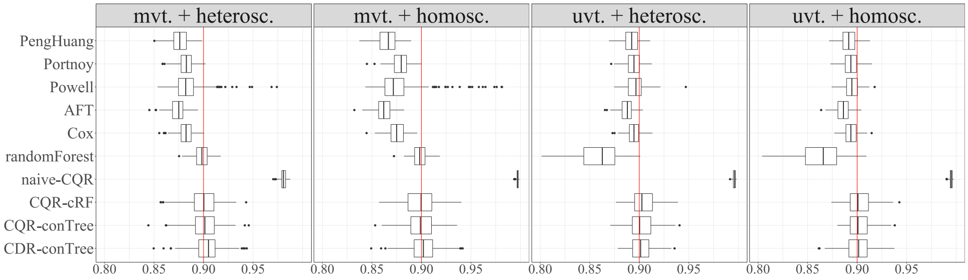

Figure 1. Empirical 90% coverage of the uncensored survival time T. 'CQR-cRF' is short for the CQR-LPB with censored quantile regression forest; 'CQR-conTree' and 'CDR-conTree' are short for the CQR- and CDR-LPB with distribution boosting. The other abbreviations are the same as in Table 1.

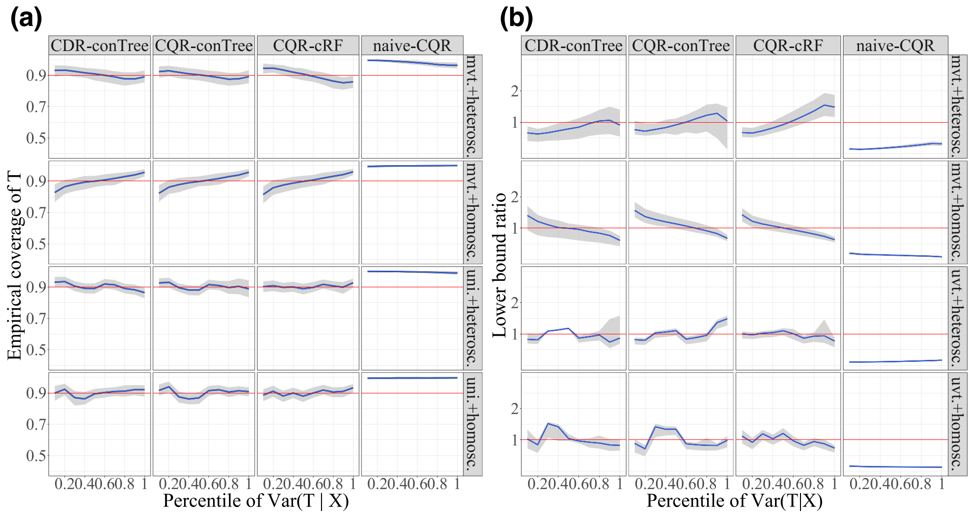

Figure 2. Results from the experiments detailed in Table 1: (a) empirical 90% conditional coverage and (b) ratio between the LPB and the theoretical quantile as a function of Var(T | X). The blue curves correspond to the mean coverage in (a) and the median ratio in (b). The grey confidence bands correspond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

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:

- *Synthetic* C: we take the censored survival time $\tilde{T}$ 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 $(X, C_{\text{syn}}, \tilde{T} \wedge C_{\text{syn}})$, and we wish to construct LPBs on $\tilde{T}$.
- *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

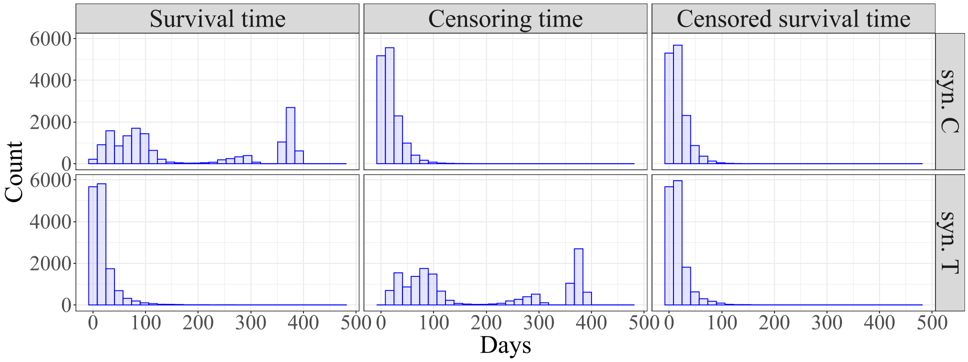

Figure 3. Histograms of the survival time, censoring time, and censored survival time defined as the minimum between the two, in each simulation setting.

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.

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

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

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 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.

6 Discussion and extensions

6.1 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; see also Korn (1986) and Schemper and Smith (1996) for an account of the two types of censoring. 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 (e.g., the patient's returning visit), C_{loss} is observable if all patients are tracked until the end of the study. 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.

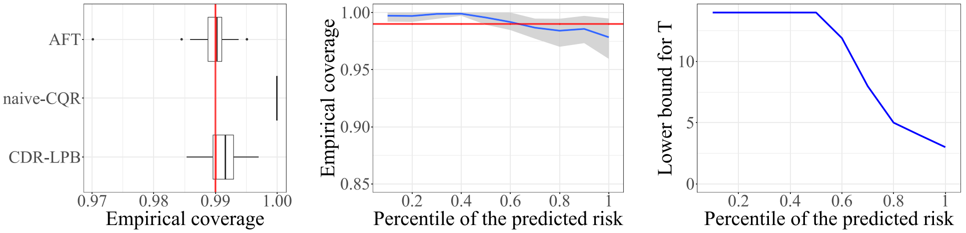

Figure 4. Results for synthetic censoring times across 100 replications: empirical coverage (left), empirical conditional coverage of the CDR-LPB (middle), and CDR-LPB as a function of the percentile of the predicted risk (right). The target coverage level is 99%. The blue curves correspond to the mean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

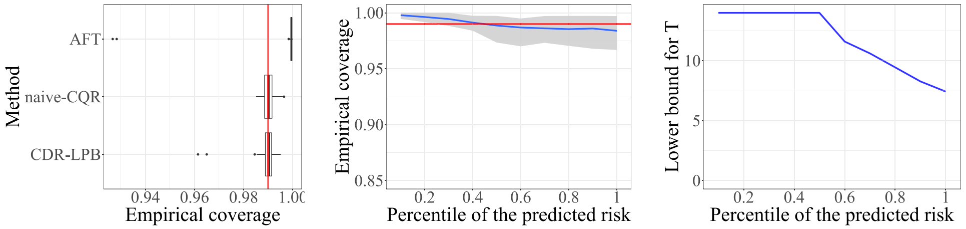

Figure 5. Results for synthetic survival times: everything else is as in Figure 4.

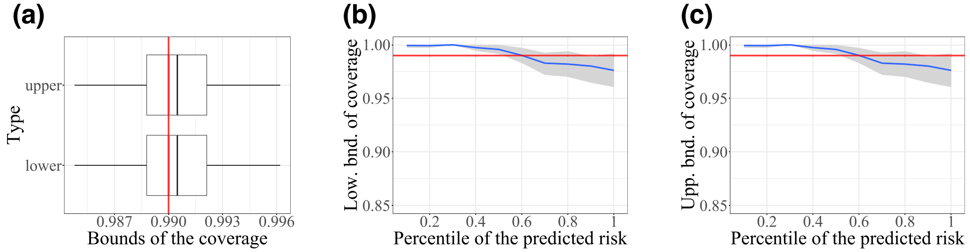

Figure 6. Analysis of the UK Biobank COVID-19 data: (a) lower and upper bounds of the empirical coverage; (b) lower and (c) upper bounds of empirical coverage as a function of the predicted risk. The target coverage level is 99%. The blue curves correspond to the mean coverage, and the grey confidence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

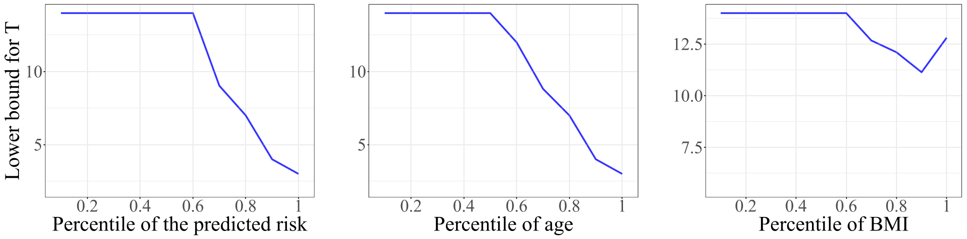

Figure 7. Analysis of the UK Biobank COVID-19 data: LPBs on the survival time of COVID-19 patients as a function of the percentiles of predicted risk (left), age (middle) and BMI (right). The target coverage level is 99%. The blue curves correspond to the median LPB across 100 sample splits.

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 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:

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

Unlike Korn (1986) and Schemper and Smith (1996), (7) does not impose any restrictions on the dependence between T and C_{loss}, which is harder to conceptualize. The assumption (7) tends to be plausible, especially when the total length of follow-up is short, 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. There are certain cases where (7) could be violated. For example, if new treatments become available during the course of a study, subjects who enter later are different from those who enter earlier as they could have been given the alternative treatments, but were not.

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 analogue 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

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

In Section D.5 of the Supplementary material, we provide an additional simulation illustrating the result of our method in this setting. 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 itself. 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 in this context.

6.2 Sharper coverage criteria

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

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

which states that $\hat{L}(X)$ is a conditionally calibrated LPB. Clearly, (8) implies valid marginal coverage. Theorems 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 5. 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,

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

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

Theorem 5 implies that no non-trivial 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 (Barber et al., 2019a; Vovk, 2012).

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 (Barber et al., 2019a; Guan, 2019; J. Lei et al., 2013; Romano, Barber, et al., 2019; Vovk, 2012). Given a partition of the covariate space $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$, Mondrian conformal inference guarantees that

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

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

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

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, Barber, et al., 2019).

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

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

The covariate shift between the observed and target distributions is

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

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.4 in Section B in the Supplementary material.

6.3 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, L. Lei and Candès (2021) proposed a method based on weighted conformal inference for counterfactual prediction under the potential outcome framework (Neyman, 1923/1990; 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

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

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:

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

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

$$w(x) = \frac{\mathrm{d}P_{X|C\geq c_0, W=1}}{\mathrm{d}P_X}(x) \propto \frac{1}{\mathbb{P}(C \geq c_0, W = 1 \mid X = x)}.$$

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 & 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.

Acknowledgments

The authors are grateful to Steven Goodman, Lucas Janson, Ying Jin, Yan Min, Chiara Sabatti, Matteo Sesia, Lu Tian and Steve Yadlowsky for their constructive feedback. This research has been conducted using the UK Biobank Resource under Application Number 27837.

Supplementary material

Supplementary data is available online at *Journal of the Royal Statistical Society* online.

Conflict of interests

None declared.

Funding

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.

Data availability

The code for reproducing the simulations in Section 4 are publicly available at https://github.com/zhimeir/cfsurv˙paper. The data used in Section 5 is provided by UK Biobank, and cannot be shared due to the privacy of the participants. An R-package to implement the procedures proposed in this paper can be found at https://github.com/zhimeir/cfsurvival.

References

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

Allison P. D. (1984). *Event history analysis: Regression for longitudinal event data* (No. 46). SAGE.

Anisimov V. V., & Fedorov V. V. (2007). Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine*, 26(27), 4958–4975. http://doi.org/10.1002/(ISSN)1097-0258

Athey S., Tibshirani J., & Wager S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178. http://doi.org/10.1214/18-AOS1709

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

Barber R. F., Candès E. J., Ramdas A., & Tibshirani R. J. (2019a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2), 455–482. https://doi.org/10.1093/imaiai/iaaa017

Barber R. F., Candes E. J., Ramdas A., & Tibshirani R. J. (2021). Predictive inference with the Jackknife+. *The Annals of Statistics*, 49(1), 486–507. https://doi.org/10.1214/20-AOS1965

Barnard K. D., Dent L., & Cook A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC Medical Research Methodology*, 10(1), 1–8. http://doi.org/10.1186/1471-2288-10-63

Breslow N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 43(1), 45–57. https://doi.org/10.2307/1402659

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. http://doi.org/10.1038/s41586-018-0579-z

Carter R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, 25(5), 429–436. http://doi.org/10.1016/j.cct.2004.07.002

Carter R. E., Sonne S. C., & 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. http://doi.org/10.1186/1471-2288-5-1

Cauchois M., Gupta S., & Duchi J. (2021). Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81), 1–42. http://jmlr.org/papers/v22/20-753.html

Chernozhukov V., Wüthrich K., & Zhu Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118. doi: 10.1073/pnas.2107794118

Cox D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

D'Amour A., Ding P., Feller A., Lei L., & Sekhon J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2), 644–654. http://doi.org/10.1016/j.jeconom.2019.10.014

Efron B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1), 1–26. http://doi.org/10.1214/aos/1176344552

Efron B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, 88(S1), S28–S59. https://doi.org/10.1111/insr.12409

Efron B., & 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., & Phillips J. (2020). Fair allocation of scarce medical resources in the time of COVID-19. *The New England Journal of Medicine*, 382(21), 2049–2055. http://doi.org/10.1056/NEJMsb2005114

Faraggi D., & Simon R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14(1), 73–82. http://doi.org/10.1002/(ISSN)1097-0258

Friedman J., & Narasimhan B. (2020). *conTree: Contrast trees and boosting*. R package version 0.2-8. https://jhfhub.github.io/conTree˙tutorial.

Friedman J. H. (2020). Contrast trees and distribution boosting. *Proceedings of the National Academy of Sciences*, 117(35), 21175–21184. http://doi.org/10.1073/pnas.1921562117

Gajewski B. J., Simon S. D., & Carlson S. E. (2008). Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in Medicine*, 27(13), 2328–2340. http://doi.org/10.1002/(ISSN)1097-0258

Geisser S. (1993). *Predictive inference* (Vol. 55). CRC Press.

Goeman J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1), 70–84. http://doi.org/10.1002/bimj.200900028

Guan L. (2022). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*. https://doi.org/10.1093/biomet/asac040

Gui J., & 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. http://doi.org/10.1093/bioinformatics/bti422

Gupta C., Kuchibhotla A. K., & Ramdas A. K. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127.

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

Hong H., & Tamer E. (2003). Inference in censored models with endogenous regressors. *Econometrica*, 71(3), 905–932. http://doi.org/10.1111/ecta.2003.71.issue-3

Hothorn T., Bühlmann P., Dudoit S., Molinaro A., & Van Der Laan M. J. (2006). Survival ensembles. *Biostatistics*, 7(3), 355–373. http://doi.org/10.1093/biostatistics/kxj011

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. http://doi.org/10.1214/08-AOAS169

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

Kaplan E. L., & Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282), 457–481. http://doi.org/10.1080/01621459.1958.10501452

Katzman J. L., Shaham U., Cloninger A., Bates J., Jiang T., & Kluger Y. (2018). DeepSurv: Personalized treatment recommender system using A cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 1–12. https://doi.org/10.1186/s12874-018-0482-1

Koenker R. (2020). *quantreg: Quantile regression*. R package version 5.75. https://CRAN.R-project.org/package=quantreg.

Korn E. L. (1986). Censoring distributions as a measure of follow-up in survival analysis. *Statistics in Medicine*, 5(3), 255–260. http://doi.org/10.1002/(ISSN)1097-0258

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

Lagakos S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1), 139–156. http://doi.org/10.2307/2529941

Lao J., Chen Y., Li Z.-C., Li Q., Zhang J., Liu J., & Zhai G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific Reports*, 7(1), 10353. http://doi.org/10.1038/s41598-017-10649-8

Lei J., G'Sell M., Rinaldo A., Tibshirani R. J., & Wasserman L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111. http://doi.org/10.1080/01621459.2017.1307116

Lei J., Robins J., & Wasserman L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501), 278–287. http://doi.org/10.1080/01621459.2012.751873

Lei J., & Wasserman L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 71–96. http://doi.org/10.1111/rssb.12021

Lei L., & Candès E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5), 911–938. http://doi.org/10.1111/rssb.v83.5

Leung K.-M., Elashoff R. M., & Afifi A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1), 83–104. http://doi.org/10.1146/publhealth.1997.18.issue-1

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

Murphy S., Rossini A., & van der Vaart A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439), 968–976. http://doi.org/10.1080/01621459.1997.10474051

Neyman J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472. Translated, edited by D. M. Dabrowska & T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1–51 (Annals of Agricultural Sciences). http://doi.org/10.1214/ss/1177012031

Peng L., & Huang Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482), 637–649. http://doi.org/10.1198/016214508000000355

Portnoy S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464), 1001–1012. http://doi.org/10.1198/016214503000000954

Powell J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, 32(1), 143–155. http://doi.org/10.1016/0304-4076(86)90016-3

Ranney M. L., Griffeth V., & 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. http://doi.org/10.1056/NEJMp2006141

Ratkovic M., & Tingley D. (Forthcoming). Estimation and Inference on nonlinear and heterogeneous effects. *Journal of Politics*.

Romano Y., Barber R. F., Sabatti C., & Candès E. J. (2020). With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2). doi: 10.1162/99608f92.03f00592

Romano Y., Patterson E., & Candes E. (2019). Conformalized quantile regression. In *Advances in neural information processing systems* (pp. 3543–3553). Curran Associates.

Romano Y., Sesia M., & Candès E. J. (2020). Classification with valid and adaptive coverage. In *Advances in neural information processing systems* (pp. 3581–3591). Curran Associates.

Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. http://doi.org/10.1093/biomet/70.1.41

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.701 (http://doi.org/10.1037/h0037350

Rubin D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 6(1), 34–58. http://doi.org/10.1214/aos/1176344064

Sadinle M., Lei J., & Wasserman L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525), 223–234. http://doi.org/10.1080/01621459.2017.1395341

Sant'Anna P. H. (2016). 'Program evaluation with right-censored data', arXiv, arXiv:1604.02642, preprint: not peer reviewed.

Saunders C., Gammerman A., & Vovk V. (1999). Transduction with confidence and credibility. In *Proceedings of the sixteenth international joint conference on artificial intelligence* (pp. 722–726). The IJCAI organization.

Scharfstein D. O., & Robins J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3), 617–634. http://doi.org/10.1093/biomet/89.3.617

Schemper M., & Smith T. L. (1996). A note on quantifying follow-up in studies of failure time. *Controlled clinical trials*, 17(4), 343–346. http://doi.org/10.1016/0197-2456(96)00075-X

Sesia M., & Candès E. J. (2020). A comparison of some conformal quantile regression methods. *Stat*, 9(1), e261. http://doi.org/10.1002/sta4.v9.1

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

Simon N., Friedman J., Hastie T., & Tibshirani R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1–3. http://doi.org/10.18637/jss.v039.i05

Stine R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392), 1026–1031. http://doi.org/10.1080/01621459.1985.10478220

Therneau T. M. (2020). *A package for survival analysis in R*. R package version 3.2-7. https://CRAN.R-project.org/package=survival.

Tibshirani R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4), 385–395. http://doi.org/10.1002/(ISSN)1097-0258

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

Vergano M., Bertolini G., Giannini A., Gristina G., Livigni S., Mistraletti G., Riccioni L., & 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(1), 165. http://doi.org/10.1186/s13054-020-02891-w

Verweij P. J., & Van Houwelingen H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24), 2305–2314. http://doi.org/10.1002/(ISSN)1097-0258

Vovk V. (2002). On-line confidence machines are well-calibrated. In *Proceedings of the 43rd annual IEEE symposium on foundations of computer science, 2002* (pp. 187–196). IEEE.

Vovk V. (2012). Conditional validity of inductive conformal predictors. In S. C. H. Hoi, & W. Buntine (Eds.), *Asian conference on machine learning* (pp. 475–490). PMLR.

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

Vovk V., Lindsay D., Nouretdinov I., & Gammerman A. (2003). *Mondrian confidence machine* (Technical Report). Royal Holloway University of London.

Wald A. (1943). An extension of Wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1), 45–55. http://doi.org/10.1214/aoms/1177731491

Wang P., Li Y., & Reddy C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6), 1–36. http://doi.org/10.1145/3214306

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. http://doi.org/10.1002/(ISSN)1097-0258

Wilks S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1), 91–96. http://doi.org/10.1214/aoms/1177731788

Witten D. M., & Tibshirani R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1), 29–51. http://doi.org/10.1177/0962280209105024

Wu M. C., & Carroll R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1), 175–188. http://doi.org/10.2307/2531905

Yang Y., & Kuchibhotla A. K. (2021). 'Finite-sample efficient conformal prediction', arXiv, arXiv:2104.13871, preprint: not peer reviewed.

Zhang H. H., & Lu W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3), 691–703. http://doi.org/10.1093/biomet/asm037