Biometrika (2024), **111**, 2, *pp.* 459–477 https://doi.org/10.1093/biomet/asad076
Advance Access publication 4 December 2023

Conformalized survival analysis with adaptive cut-offs

By YU GUI, ROHAN HORE

Department of Statistics, University of Chicago,
5747 South Ellis Avenue, Chicago, Illinois 60637, U.S.A.
yugui@uchicago.edu rohanhere@uchicago.edu

ZHIMEI REN

Department of Statistics and Data Science, The Wharton School,
University of Pennsylvania, 265 South 37th Street, Philadelphia, Pennsylvania 19104, U.S.A.
zren@wharton.upenn.edu

and RINA FOYGEL BARBER

Department of Statistics, University of Chicago,
5747 South Ellis Avenue, Chicago, Illinois 60637, U.S.A.
rina@uchicago.edu

Summary

This paper introduces an assumption-lean method that constructs valid and efficient lower predictive bounds for survival times with censored data. We build on recent work by Candès et al. (2023), whose approach first subsets the data to discard any data points with early censoring times and then uses a reweighting technique, namely, weighted conformal inference (Tibshirani et al., 2019), to correct for the distribution shift introduced by this subsetting procedure. For our new method, instead of constraining to a fixed threshold for the censoring time when subsetting the data, we allow for a covariate-dependent and data-adaptive subsetting step, which is better able to capture the heterogeneity of the censoring mechanism. As a result, our method can lead to lower predictive bounds that are less conservative and give more accurate information. We show that in the Type-I right-censoring setting, if either the censoring mechanism or the conditional quantile of the survival time is well estimated, our proposed procedure achieves nearly exact marginal coverage, where in the latter case we additionally have approximate conditional coverage. We evaluate the validity and efficiency of our proposed algorithm in numerical experiments, illustrating its advantage when compared with other competing methods. Finally, our method is applied to a real dataset to generate lower predictive bounds for users' active times on a mobile app.

Some key words: Censoring; Conformal inference; Prediction interval; Random forest; Survival time.

1. Introduction

1.1. *Background*

Survival analysis lies at the core of many important questions in clinical trials (Fleming & Lin, 2000, Singh & Mukhopadhyay, 2011), ecology (Muenchow, 1986) and other

applied fields. In particular, one important problem is that of studying the behaviour of the survival time T, and how it relates to other features of the data, which we denote by a potentially high-dimensional feature vector X. Modelling the association between X and T can in turn play a crucial role in enabling more useful and reliable policy making. The major challenge is that these survival times are only partially observed due to censoring (Leung et al., 1997), which makes the statistical analysis quite nonroutine; we are only able to observe the survival time T if it occurs no later than some censoring time C. For example, T may be the survival time of a patient, measured as time since diagnosis, which may be censored at a time C that denotes the endpoint of the study that follows the patient.

One of many goals of survival analysis is to infer the survival function, which is the probability of survival beyond a given time, given the censored data. The Kaplan–Meier curve (Kaplan & Meier, 1958) can produce such inferences for subpopulations with a particular covariate structure while making no assumption on the distribution of survival times, but it requires sufficiently many events in each subgroup (Kalbfleisch & Prentice, 2011). This assumption is no longer realistic in the modern era of big data, where, with the ever-increasing ability to collect and store data, we can have access to a large number of potentially continuous covariates.

Over the years, many tools have been developed to cope with such high dimensionality, offering estimation of the conditional survival function. One popular example is the Cox model that posits a proportional hazard model: an unspecified nonparametric baseline is modified via a parametric model describing how the hazard varies in response to explanatory covariates (Cox, 1972, Breslow, 1975). Other popular parametric approaches include the accelerated failure time model (Cox, 1972, Wei, 1992) and the proportional odds model (Murphy et al., 1997, Harrell, 2015). More recently, we have witnessed more complex survival analysis methods that are based on machine learning/deep learning (Faraggi & Simon, 1995, Tibshirani, 1997, Gui & Li, 2005, Katzman et al., 2018, Lao et al., 2017, Wang et al., 2019, Li & Bradic, 2020). Despite the success of these methods in many areas, it remains largely unclear how to provide reliable uncertainty quantification for these methods. This is mainly because they posit model assumptions that are hard to verify and/or the algorithms themselves are too complicated to be analysed. For these reasons, it is desirable to find a more assumption-lean or distribution-free approach towards reliable inference in survival analysis.

The recent work of Candès et al. (2023) proposes such an approach, which we describe in detail below. As the target of inference, they proposed computing a $100(1-\alpha)\%$ lower prediction bound, LPB, for the survival time of a patient/unit, where α is a prespecified level; it means that the patient/unit is expected to survive beyond this predicted time with at least $100(1-\alpha)\%$ probability. The LPB is used to provide a summary of what we can infer about the individual's survival time given available data. The LPB can be low when either the true survival time is low or there is not enough information for us to get an informative lower bound; in other words, insufficient data should not lead to an invalid claim, but instead may lead to a less informative output.

1.2. *Defining the lower prediction bound*

Let $X \in \mathcal{X}$ denote the covariate vector, $T \in \mathbb{R}_{\geqslant 0}$ the survival time and $C \in \mathbb{R}_{\geqslant 0}$ the censoring time. Under censoring, the survival time T is observed only if it occurs before the censoring time C. In other words, while the features X and the censoring time C are

both observed, the survival time is observed only indirectly, via the censored survival time as $\tilde{T} = \min(T, C)$, which may not be equal to T.

We now give the formal definition of a marginally calibrated LPB. Throughout, for a joint distribution P on (X, C, T), we write P_X, $P_{(X,T)}$, $P_{(X,\tilde{T})}$, etc. to denote the corresponding marginal distributions, and $P_{C|X}$, $P_{T|X}$, $P_{\tilde{T}|X}$, etc. to denote the corresponding conditional distributions.

Definition 1 (Marginally calibrated LPB). *Let* $(X_i, C_i, T_i) \overset{\text{i.i.d.}}{\sim} P$ *for data points* $i = 1, \ldots, n$, *and let* $\hat{L}$ *be a function of the observed data* $\mathcal{D} = \{(X_i, C_i, \tilde{T}_i) : 1 \leqslant i \leqslant n\}$, *where* $\tilde{T}_i = \min(T_i, C_i)$ *is the censored survival time. Then we say that* $\hat{L}$ *is a marginally calibrated LPB at level* $1 - \alpha$ *if it satisfies*

$$\mathbb{P}_{(X,T)\sim P}\{T \geqslant \hat{L}(X)\} \geqslant 1 - \alpha, \tag{1}$$

where this probability is taken with respect to both the available data $\mathcal{D}$ *and a new data point* $(X, T) \sim P_{(X,T)}$.

The marginally calibrated LPB provides a guarantee in an average sense; that is, over all the possible draws of the data, the coverage of the LPBs is guaranteed. However, in practical settings, we may be more concerned about the coverage guarantee we can obtain given the data at hand. In such settings, the probably-approximately-correct, PAC, type LPB defined below can be more informative, see also Vovk (2012), Bates et al. (2021), Angelopoulos et al. (2022) and Jin et al. (2022).

Definition 2 (PAC-type LPB). *Under the same notation as in Definition 1, we say that* $\hat{L}$ *is a PAC-type LPB at level* α *with tolerance* δ *if, with probability at least* $1 - \delta$ *over the draw of* $\mathcal{D}$,

$$\mathbb{P}_{(X,T)\sim P}\{T \geqslant \hat{L}(X) \mid \mathcal{D}\} \geqslant 1 - \alpha,$$

where the probability is now taken with respect to a new data point $(X, T) \sim P_{(X,T)}$.

Throughout, we adopt the following conditionally independent censoring assumption.

Assumption 1 (*Conditionally independent censoring*). The joint distribution P of (X, C, T) satisfies $C \perp\!\!\!\perp T \mid X$.

This assumption is standard in the survival analysis literature, in order to ensure identifiability of the problem; see, e.g., Kalbfleisch & Prentice (2011).

1.3. *An initial approach: inference on the censored survival time*

As discussed by Candès et al. (2023), since the censored survival time $\tilde{T}$ cannot be larger than T by definition, any valid lower bound on $\tilde{T}$ is trivially a lower bound on T. In other words, if an estimated lower bound $\hat{L}$ satisfies $\mathbb{P}_{(X,C,T)\sim P}\{\tilde{T} \geqslant \hat{L}(X)\} \geqslant 1 - \alpha$ then, trivially, Definition 1 is satisfied and so $\hat{L}$ is a marginally calibrated LPB. Similarly, if $\mathbb{P}\{\tilde{T} \geqslant \hat{L}(X) \mid \mathcal{D}\} \geqslant 1 - \alpha$ with probability at least $1 - \delta$ then, by Definition 2, $\hat{L}$ is a PAC-type LPB. Since the censored survival time $\tilde{T}$ can be observed in the dataset at hand, and so $\hat{L}$ can be constructed to satisfy this property, this provides a mechanism for providing a valid LPB.

However, if the censoring time C is often substantially smaller than T then a valid lower bound on $\tilde{T}$ may be extremely conservative as a lower bound on T itself, thus reducing the utility of the constructed LPB. This suggests that such an approach may not be optimal for most applications. On the other hand, Candès et al. (2023) proved that, in the absence of any assumptions on the distribution P on (X, C, T), it is impossible to improve on this type of approach; specifically, their result (Candès et al., 2023, Theorem 1) proves that, for any construction $\hat{L}$ that satisfies Definition 1 universally over all distributions P, $\hat{L}$ must also satisfy $\mathbb{P}\{\tilde{T} \geqslant \hat{L}(X)\} \geqslant 1 - \alpha$. This motivates their introduction of an additional assumption, as we describe next.

1.4. *The approach of Candès et al. (2023): a cut-off on the censoring time*

As described above, constructing an LPB on the censored survival time $\tilde{T}$ may be too conservative in applications where the censoring time C is frequently low, leading to censored times $\tilde{T}$ that are far smaller than the true target of inference T. The approach of Candès et al. (2023) is to avoid this issue by discarding any training data points where C is very low; specifically, for a constant cut-off c_0, they subset the data $\mathcal{D}$ to keep only data points $(X_i, C_i, \tilde{T}_i)$ for which $C_i \geqslant c_0$. After this filtering step, any lower bound $\hat{L}$ on the remaining censored survival time $\tilde{T}$ is no longer necessarily overly conservative, since the condition $C \geqslant c_0$, with a well chosen c_0, ensures that $\tilde{T}$ is less likely to be far smaller than T. Thus, we can proceed by constructing an LBP $\hat{L}$ that is a lower bound on $\tilde{T}$ in this new training sample.

Of course, we must then be careful about biasing the results because of this cut-off. In particular, since the event $C \geqslant c_0$ may be highly dependent on the covariates X, the remaining data are drawn from a distribution that is different from the target distribution P. To be more precise, writing $P^{\geqslant c_0}$ to denote the distribution of a data point $(X, C, T) \sim P$ given the event $C \geqslant c_0$, we see that the remaining data consist of samples from $P^{\geqslant c_0}$, while the inference goal is to provide coverage under the original distribution P. In other words, we would like to ensure that the marginal coverage bound (1) holds, but calibrating $\hat{L}(\cdot)$ naïvely on the remaining data would instead only ensure that $\mathbb{P}_{(X,T)\sim P^{\geqslant c_0}}\{T \geqslant \hat{L}(X)\} \geqslant 1 - \alpha$, or, equivalently, $\mathbb{P}_{(X,T,C)\sim P}\{T \geqslant \hat{L}(X) \mid C \geqslant c_0\} \geqslant 1 - \alpha$.

To account for this shift in the distribution, Candès et al. (2023) utilized the method of conformal prediction under covariate shift (Tibshirani et al., 2019), which builds on the well-known conformal prediction framework for distribution-free predictive inference (Vovk et al., 2005). To do so, they additionally assumed that we have exact or approximate knowledge of the dependence of censoring time C on the covariates X, that is, knowledge of $P_{C|X}$ or, more specifically, $\mathbb{P}(C \geqslant c_0 \mid X)$. With this additional information, we can reweight the remaining data points to correct for the change in distribution; essentially, similarly to inverse propensity score weighting, weights $1/\mathbb{P}_P(C \geqslant c_0 \mid X)$ can account for the difference between the target distribution P and its filtered version $P^{\geqslant c_0}$. Of course, the best value of c_0 will depend on the data distribution, and in practice can be chosen on a training set.

1.5. *Our approach: the benefits of a covariate-adaptive cut-off*

In the method described above, how should the cut-off c_0 be chosen? The choice of c_0 presents a trade-off: if c_0 is chosen to be too small then the inequality $\tilde{T} \leqslant T$ might be quite loose, and the constructed LPB $\hat{L}$ might still be very conservative even after filtering the data with the cut-off. On the other hand, if c_0 is chosen to be too large then $\mathbb{P}_P(C \geqslant c_0 \mid X)$ may be quite small, at least, for many values of X, leading to a low effective sample size,

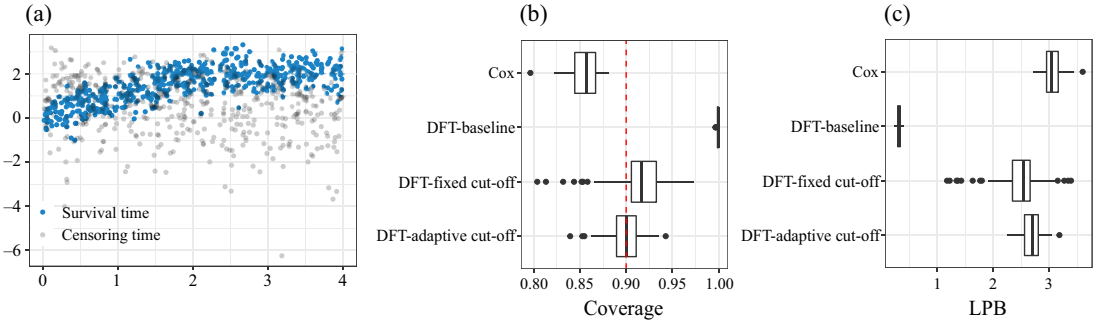

Fig. 1. (a) An illustration of the training sample for one trial of the experiment. (b) Boxplot of the coverage rate; the red dashed line corresponds to the target coverage rate $1-\alpha = 90\%$. (c) Boxplot of the LPBs. The results are from 100 independent trials.

large weights $1/\mathbb{P}_P(C \geqslant c_0 \mid X)$ on these data points and highly unstable behaviour. In fact, it is not always possible to find a constant c_0 that yields good LPBs, especially in cases when the censoring time varies substantially with respect to the covariates X. Selecting a large value of c_0 could cause highly unstable LPBs in areas where censoring times are low, whereas selecting a small value of c_0 leads to conservative LPBs in areas where censoring times are actually high. To be more specific, think of a simple example where $X \sim \mathrm{Un}([0,1])$ and $C = a\mathbb{1}\{X \geqslant \frac{1}{2}\} + b\mathbb{1}\{X < \frac{1}{2}\}$ with $a \gg b$; choosing c_0 to be greater than b requires dropping half of the data and leads to increased variability; instead, selecting a $c_0 \leqslant b$ yields very conservative LPBs for $X \geqslant \frac{1}{2}$.

From the above discussion, we can see that it may be beneficial to allow c_0 to depend on X. That is, if $\mathbb{P}_P(C \geqslant c_0 \mid X)$ is extremely small then we may instead need to choose a lower value of c_0 to avoid high variance, but if $\mathbb{P}_P(C \geqslant c_0 \mid X)$ is close to 1 then we can afford to increase the value of c_0, thus avoiding an overly conservative LPB. To illustrate the benefits of this more flexible approach, we show a small simulated example.

We consider a univariate-covariate case, where T and C depend on X via different models, the details are given in § 4 below. Figure 1(a) visualizes the censoring time and survival time as functions of the covariate. In this example, units with larger values of X tend to have lower censoring times ($P_{C|X} = \mathrm{Ex}\{0.25 + (6+x)/100\}$), and thus we should choose a lower value of c_0 to avoid high variance, i.e., to avoid overly large weights $1/\mathbb{P}_P(C \geqslant c_0 \mid X)$; units with smaller values of X, on the other hand, tend to have larger values of C and so we can afford to increase the value of c_0, leading to a less conservative LPB.

From this model, $n = 2000$ independent samples are generated. We compare (i) the baseline method introduced in § 1.3, referred to as DFT-baseline, where DFT is short for distribution-free LPB for T, (ii) the fixed cut-off method of Candès et al. (2023), referred to as DFT-fixed cut-off, (iii) our new adaptive cut-off method, referred to as DFT-adaptive cut-off and to be defined shortly and (iv) the Cox parametric model. The generated LPBs are then evaluated with an independent dataset of 5000 test samples, and we display the coverage rate and the resulting LPB in Figs. 1(b) and 1(c), respectively, with results gathered from 100 independent trials. The parametric method fails to cover the true survival time with desired probability; the baseline method and, to a lesser extent, the fixed cut-off method are conservative in this setting, returning a low, i.e., less informative, LPB. On the other hand, our adaptive cut-off method is able to avoid undercoverage or overcoverage; it essentially achieves the target coverage rate and returns a higher, i.e., more precise, LPB.

2. Background

2.1. *Covering the censored survival time via conformalized quantile regression*

As described in § 1.3, it is possible to provide an LPB on T with no further assumptions by simply finding a lower bound on the censored survival time $\tilde{T} \leqslant T$. To do so, one approach is to use the conformalized quantile regression framework of Romano et al. (2019). To begin, we first partition the available n data points into two datasets: a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$; for instance, we partition into two sets of size $n/2$. Using the training set, we fit a quantile regression, $x \mapsto \hat{q}_\alpha(x)$, which estimates the conditional α-quantile of T given X. This may be done using an arbitrary algorithm, for instance, linear regression or random forests. If this quantile regression were fitted accurately then we could simply use $\hat{q}_\alpha(X)$ as an LPB for T. If indeed this is the α-quantile of $T \mid X$ then $T \geqslant \hat{q}_\alpha(X)$ holds with probability $1-\alpha$, as desired. However, due to potential issues of overfitting, model misspecification, etc., we cannot rely on this being the case, and so the calibration set is then used to correct for any errors in the initial model-fitting stage. For each calibration point $i \in \mathcal{I}_2$, define a score $V_i = \hat{q}_\alpha(X_i) - \tilde{T}_i$, and then define the LPB as

$$\hat{L}_{\text{baseline}}(X) = \hat{q}_\alpha(X) - Q_{1-\alpha}\Big(\sum_{i\in\mathcal{I}_2} \frac{1}{1+|\mathcal{I}_2|}\delta_{V_i} + \frac{1}{1+|\mathcal{I}_2|}\delta_{+\infty}\Big),$$

where $Q_{1-\alpha}(\cdot)$ denotes the $(1-\alpha)$-quantile of a distribution and δ_v is the point mass at v. The intuition here is that the $Q_{1-\alpha}(\cdot)$ term adds a correction to the original fitted model to ensure that $\hat{L}_{\text{baseline}}(X)$ has the right coverage level on the calibration set drawn independent and identically distributed from P, and will thus have the right coverage level on a future draw (X, T) from $P_{(X,T)}$ as well.

The resulting value $\hat{L}_{\text{baseline}}(X)$ may be higher (less conservative) or lower (more conservative) than the initial fitted model $\hat{q}_\alpha(X)$, depending on whether the original fitted model $\hat{q}_\alpha$ is over- or under-covering on the calibration set. In practice, it is likely that we will have undercoverage of the original fitted model (see, e.g., the simulation results of Romano et al. (2019), Lei & Candès (2021), Candès et al. (2023) and Fig. 6 in § 5 below, leading to a quantile $Q_{1-\alpha}(\cdot)$ that is positive, and an LPB $\hat{L}(X)$ that is lower, i.e., more conservative, than the original fitted model.

The following result proves that this is a marginally calibrated LPB.

Theorem 1 (Adapted from Theorem 1 of Romano et al., 2019). *Suppose that* $(X_i, C_i, T_i) \overset{\text{i.i.d.}}{\sim} P$. *Then* $\hat{L}_{\text{baseline}}(X)$ *is a marginally calibrated LPB at level* $1-\alpha$ *and, moreover, satisfies*

$$\mathbb{P}_{(X,T,C)\sim P}\{\tilde{T} \geqslant \hat{L}_{\text{baseline}}(X)\} \geqslant 1-\alpha.$$

Since $\tilde{T}$ may often be much smaller than T if the censoring is severe, this result indicates that such an LPB may be quite conservative as a lower bound for T. This conservativeness is however inescapable without further assumptions; Candès et al. (2023, Theorem 1) established that, under mild conditions, for any marginally calibrated LPB $\hat{L}$ for the uncensored survival time T that is valid universally over all distributions P on the data, $\hat{L}$ must also be an LPB for $\tilde{T}$ whenever $P_{(C,T)}$ is either discrete or continuous.

2.2. *Using a fixed threshold c_0*

We now give details of the proposed method of Candès et al. (2023), which uses a fixed threshold c_0 to avoid an overly conservative LPB. As mentioned above, their work shows that, without further assumptions, it is not possible to improve on the LPB for $\tilde{T}$; therefore, they make the additional assumption that the conditional distribution $P_{C|X}$ is known, or is estimated accurately.

As for conformalized quantile regression, their method begins by partitioning the data into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$, and uses the training set to fit a quantile regression, $x \mapsto \hat{q}_\alpha(x)$, for the conditional α-quantile of T given X. While their proposed method is defined via a more general construction, here we focus on a single version that is most relevant for comparison to our own methods. The cut-off c_0 for the censoring time may also be chosen as a function of the training data. Furthermore, define $\hat{w}(x)$ to be an estimate of $1/\mathbb{P}(C \geqslant c_0 \mid X = x)$ or, approximately proportional to this quantity, also fitted on the training data.

Next, on the calibration set, we use c_0 to filter the data and define $\mathcal{I}_2' = \{i \in \mathcal{I}_2 \colon C_i \geqslant c_0\}$. For all these remaining calibration points, note that $\tilde{T}_i \wedge c_0 = T_i \wedge c_0$, that is, $T_i \wedge c_0$ is observed. We then calculate scores $V_i = \hat{q}_\alpha(X_i) - T_i \wedge c_0$ for all $i \in \mathcal{I}_2'$, and return the LPB

$$\hat{L}_{\text{fixed cut-off}}(X) = \hat{q}_\alpha(X) - Q_{1-\alpha}\left(\frac{\sum_{i\in\mathcal{I}_2'} \hat{w}(X_i)\delta_{V_i} + \hat{w}(X)\delta_{+\infty}}{\sum_{i\in\mathcal{I}_2'} \hat{w}(X_i) + \hat{w}(X)}\right).$$

The intuition here is that the calibration set $\mathcal{I}_2'$ consists of data points drawn from the shifted distribution $P^{\geqslant c_0}$, and the likelihood ratio between the target distribution $\mathbb{P}$ and this distribution $\mathbb{P}^{\geqslant c_0}$ is $\mathbb{P}(C \geqslant c_0)/\mathbb{P}(C \geqslant c_0 \mid X)$; since $\hat{w}(X)$ is an estimate of the likelihood ratio, up to constants, reweighting the calibration data points with weights $\hat{w}(X_i)$ ensures coverage with respect to the actual target distribution $\mathbb{P}$.

Building on the framework of conformal prediction with covariate shift (Tibshirani et al., 2019), the result of Candès et al. (2023) proves that this construction yields a valid LPB.

THEOREM 2 (PROPOSITION 1 OF CANDÈS ET AL., 2023). *Suppose that* $(X_i, C_i, T_i) \overset{\text{i.i.d.}}{\sim} P$ *and that* $\hat{w}(x) = 1/\mathbb{P}(C \geqslant c_0 \mid X = x)$, *i.e., this probability was fitted exactly. Then* $\hat{L}_{\text{fixed cut-off}}(X)$ *is a marginally calibrated LPB for* $T \wedge c_0$, *and therefore also for* T.

Moreover, Candès et al. (2023, Theorem 2) established a double robustness result: if either $\hat{w}(x)$ was fitted accurately, i.e., is a good approximation of $1/\mathbb{P}(C \geqslant c_0 \mid X = x)$, or the quantile regression was fitted accurately, i.e., $\hat{q}_\alpha(x)$ is a good approximation of the α-quantile of $T \mid X$, then $\hat{L}_{\text{fixed cut-off}}$ approximately satisfies the criterion for a marginally calibrated LPB.

3. CONFORMALIZED SURVIVAL ANALYSIS WITH ADAPTIVE CUT-OFFS

3.1. *Our procedure*

As before, we first partition the data into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$. On the training set, we fit a family of estimated quantile regression functions, $(x, a) \mapsto \hat{q}_a(x)$, mapping x to the estimated a-quantile of the conditional distribution of T given $X = x$ for all $a \in [0, 1]$. We assume that, for any x, $a \mapsto \hat{q}_a(x)$ is nondecreasing. If our estimators $\hat{q}_a$ are computed independently for each a and this constraint is violated, monotonicity can

easily be restored via sorting the outputs; see, e.g., Koenker (1994). We define $\hat{q}_a$ to be $-\infty$ at $a = 0$. In contrast, for the existing methods defined in § 2, this regression is run only at a single value of a.

Next, we need to use the calibration set in order to choose a value of a, for which returning $\hat{q}_a(X)$ is a valid LPB; that is, we need to find a value a such that $\mathbb{P}\{T < \hat{q}_a(X)\} = \alpha$, where we implicitly treat the fitted quantile function $\hat{q}_a$ as fixed, and take the probability over $(X, T) \sim P$. If the original regression were estimated perfectly then we would expect to return $a = \alpha$, i.e., the estimated α-quantile $\hat{q}_\alpha(X)$ would already be a valid LPB. In practice, as discussed earlier, we expect to see overfitting in most real-data settings and thus we expect to return $a < \alpha$. To choose a appropriately, we could consider solving for a in the expression

$$\begin{aligned}
\alpha &= \mathbb{P}\{T < \hat{q}_a(X)\} \\
&= \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\}] \\
&= \mathbb{E}\left[\mathbb{P}\{T < \hat{q}_a(X) \mid X\}\frac{\mathbb{P}\{\hat{q}_a(X) \leqslant C \mid X\}}{\mathbb{P}\{\hat{q}_a(X) \leqslant C \mid X\}}\right] \\
&\approx \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\}\mathbb{P}\{\hat{q}_a(X) \leqslant C \mid X\}\hat{w}_a(X)] \\
&= \mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \leqslant C \mid X\}\hat{w}_a(X)] \\
&= \mathbb{E}[\mathbb{1}\{T < \hat{q}_a(X) \leqslant C\}\hat{w}_a(X)], \qquad (2)
\end{aligned}$$

where now $\hat{w}_a(x)$ is chosen to be approximately equal to $1/\mathbb{P}\{C \geqslant \hat{q}_a(X) \mid X = x\}$, and where the next-to-last step holds by Assumption 1. If, instead, we only assume that our estimate $\hat{w}_a(x)$ is proportional to $1/\mathbb{P}\{C \geqslant \hat{q}_a(X) \mid X = x\}$ then we want to solve for a in the equation

$$\begin{aligned}
\alpha &= \mathbb{P}\{T < \hat{q}_a(X)\} \\
&\approx \frac{\mathbb{E}[\mathbb{P}\{T < \hat{q}_a(X) \mid X\}\mathbb{P}\{\hat{q}_a(X) \leqslant C \mid X\}\hat{w}_a(X)]}{\mathbb{E}[\mathbb{P}\{\hat{q}_a(X) \leqslant C \mid X\}\hat{w}_a(X)]} \\
&= \frac{\mathbb{E}[\mathbb{1}\{T < \hat{q}_a(X) \leqslant C\}\hat{w}_a(X)]}{\mathbb{E}[\mathbb{1}\{\hat{q}_a(X) \leqslant C\}\hat{w}_a(X)]}. \qquad (3)
\end{aligned}$$

While the events $\mathbb{1}\{T_i < \hat{q}_a(X_i)\}$ cannot be observed on the calibration set, since we only observe the censored survival time, $\tilde{T}_i$, the filtered events $\mathbb{1}\{T_i < \hat{q}_a(X_i) \leqslant C_i\}$ can be observed, since if $C_i \geqslant \hat{q}_a(X_i)$ then $\mathbb{1}\{T_i < \hat{q}_a(X_i)\} = \mathbb{1}\{\tilde{T}_i < \hat{q}_a(X_i)\}$. Therefore, the calibration set can indeed be used to find a value a so that (2) or (3) is approximately satisfied.

We now formally describe how to select a using the calibration set. For each value a, we estimate the miscoverage rate $\mathbb{P}\{T < \hat{q}_a(X)\}$ as

$$\hat{\alpha}(a) = \frac{\sum_{i\in\mathcal{I}_2} \hat{w}_a(X_i)\mathbb{1}\{T_i < \hat{q}_a(X_i) \leqslant C_i\}}{\sum_{i\in\mathcal{I}_2} \hat{w}_a(X_i)\mathbb{1}\{\hat{q}_a(X_i) \leqslant C_i\}}.$$

This empirical quantity estimates $\alpha^*(a) = \mathbb{P}\{T < \hat{q}_a(X)\}$. Since our aim is to find a value of a sufficiently small so that $\alpha^*(a) \leqslant \alpha$, we instead search for a satisfying $\hat{\alpha}(a) \leqslant \alpha$. However, while $\alpha^*(a)$ is monotone in a, since $\hat{q}_a(x)$ is monotone in a, this property may not hold for the estimator $\hat{\alpha}(a)$; we therefore define $\hat{a} = \sup\{a \in [0, 1]\colon \sup_{a' \leqslant a} \hat{\alpha}(a') \leqslant \alpha\}$. Finally, we output the LPB $\hat{L}(X) := \hat{q}_{\hat{a}}(X)$.

Below, we give a double robustness result proving that this choice of $\hat{L}$ is approximately a marginally calibrated LPB, as long as either the weights $\hat{w}_a$ or the quantiles $\hat{q}_a$ are fitted accurately; furthermore, when the quantiles are fitted accurately, the LPBs are approximately conditionally valid. Before giving our theoretical results, we first present a more general form of this procedure.

3.2. *A generalized procedure*

The procedure described above tends to perform well in settings where $\hat{q}_a(x)$, for relevant values of a, is not too large, so that $\mathbb{P}\{C \geqslant \hat{q}_a(x) \mid X = x\}$ is not close to zero and the weights $\hat{w}_a(X_i)$ on the calibration points are not too large. In other settings, however, the procedure may be somewhat unstable. Specifically, in scenarios where C is often much smaller than T, as in the example given in § 1.5, we might have a very small probability $\mathbb{P}\{C \geqslant \hat{q}_a(X) \mid X = x\}$; this is problematic since the inverse weight $\hat{w}(x)$ will then be extremely large. To alleviate this, we now generalize the procedure sketched above to allow for a more stable and robust method. Define a family of functions $(x, a) \mapsto \hat{f}_a(x)$ for $x \in \mathcal{X}$ and $a \in \mathcal{A}$ that are fitted on the training set such that, for each fixed x, this map is nondecreasing in a, note that the estimated quantiles, $\hat{q}_a(x)$, are simply a special case. Our aim is now to use the calibration set in order to choose a so that $\hat{L}(X) = \hat{f}_a(X)$ offers a calibrated LPB. With the same rationale as before, we wish to find a to satisfy

$$\begin{aligned}\alpha &= \mathbb{P}\{T < \hat{f}_a(X)\} \\ &\approx \frac{\mathbb{E}[\mathbb{P}\{T < \hat{f}_a(X) \mid X\}\mathbb{P}\{\hat{f}_a(X) \leqslant C \mid X\}\hat{w}_a(X)]}{\mathbb{E}[\mathbb{P}\{\hat{f}_a(X) \leqslant C \mid X\}\hat{w}_a(X)]} \\ &= \frac{\mathbb{E}[\mathbb{1}\{T < \hat{f}_a(X) \leqslant C\}\hat{w}_a(X)]}{\mathbb{E}[\mathbb{1}\{\hat{f}_a(X) \leqslant C\}\hat{w}_a(X)]},\end{aligned}$$

where $\hat{w}_a$ is again fitted on the training data, but is now chosen to be approximately proportional to $1/\mathbb{P}\{C \geqslant \hat{f}_a(X) \mid X = x\}$.

From this point on, we proceed exactly as before, but with $\hat{f}_a$ in place of $\hat{q}_a$; we define

$$\hat{\alpha}(a) = \frac{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i)\mathbb{1}\{T_i < \hat{f}_a(X_i) \leqslant C_i\}}{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i)\mathbb{1}\{\hat{f}_a(X_i) \leqslant C_i\}},$$

which estimates $\alpha^*(a) = \mathbb{P}\{T < \hat{f}_a(X)\}$. As before, we compute

$$\hat{a} = \sup\left\{a \in [0, 1]: \sup_{a' \leqslant a} \hat{\alpha}(a') \leqslant \alpha\right\}, \tag{4}$$

and return the LPB $\hat{L}(X) := \hat{f}_{\hat{a}}(X)$.

In this more general procedure, how should the family $\hat{f}_a(x)$ be chosen? The LPB will be approximately valid regardless of our choice, but the utility of the method will depend strongly on choosing a reasonable family of functions. We consider two goals when choosing the family.

(i) We would like to closely approximate the oracle LPB, $L(X) = q_\alpha(X)$, where $q_\alpha(x)$ is the true α-quantile of T given $X = x$. As a result, for a such that $\hat{q}_a(x)$ is close to $q_\alpha(x)$, we would like to have $\hat{f}_a(x) \approx \hat{q}_a(x)$, our estimated quantiles for T given X.

(ii) On the other hand, we would like for the weights $\hat{w}_a(x)$ to not be too large, or, equivalently, for $\mathbb{P}\{C \geqslant \hat{f}_a(X) \mid X = x\}$ to not be too small for any a. Consequently, we might want to require $\hat{f}_a(x) \leqslant \hat{q}^C_{1-\beta}(x)$, where $\hat{q}^C_{1-\beta}(x)$ estimates the conditional quantile of C given $X = x$, and we choose some constant value β.

To balance between these two goals, we propose selecting $\hat{f}_a(x) = \min\{\hat{q}_a(x), \hat{q}^C_{1-\beta}(x)\}$. As for the choice of β, we use $\beta = 1/\log|\mathcal{I}_2|$ in our implementation such that $\hat{w}_a(X_i) \leqslant \log|\mathcal{I}_2| = o(\sqrt{|\mathcal{I}_2|})$. In the simulations, we compare this choice against the canonical version of the method with $\hat{f}_a(x) = \hat{q}_a(x)$, to see how this new choice adds stability to the method.

Next we describe how the threshold $\hat{a}$ in (4) can be computed efficiently in practice. We note that $\sup_{a' \leqslant a} \hat{\alpha}(a')$ is a nondecreasing piecewise constant function in a, with no more than $2n$ knots, values of a at which the indicators $\mathbb{1}\{T_i < \hat{f}_a(X_i) \leqslant C_i\}$ or $\mathbb{1}\{\hat{f}_a(X_i) \leqslant C_i\}$ change signs. Define $\bar{a}_i = \sup_{a \in [0,1]}\{\hat{f}_a(X_i) \leqslant \tilde{T}_i\}, \tilde{a}_i = \sup_{a \in [0,1]}\{\hat{f}_a(X_i) \leqslant C_i\}$ and $\mathcal{A}_1 = \{\bar{a}_i \colon i = 1, \ldots, n\}, \mathcal{A}_2 = \{\tilde{a}_i \colon i = 1, \ldots, n\}$. Then, by definition, the breakpoints of the piecewise constant map $a \mapsto \hat{\alpha}(a)$ must all lie in $\mathcal{A}_1 \cup \mathcal{A}_2$. In the implementation, in order to obtain $\hat{a}$, we only need to search through the finite grids

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \{0\}. \tag{5}$$

A complete description of the general procedure is given in the following algorithm.

Algorithm 1. Conformalized survival analysis with adaptive cut-offs.

Input: level α; data $\mathcal{D} = (X_i, \tilde{T}_i, C_i)_{i \in [n]}$.

1: Split the data into two folds: the training fold $\mathcal{I}_1$ and the calibration fold $\mathcal{I}_2$.
2: Using $\mathcal{I}_1$ as input, apply any algorithm to fit the candidate LPBs $\{\hat{f}_a(\cdot)\}_{a \in [0,1]}$.
3: Using $\mathcal{I}_1$ as input, apply any algorithm to construct estimates $\hat{w}_a(x)$ of $\mathbb{P}\{C \geqslant \hat{f}_a(x) \mid X = x\}$.
4: Determine $\mathcal{A}$ according to (5).
5: For a in $\mathcal{A}$ do:
Compute the estimated miscoverage rate

$$\hat{\alpha}(a) = \frac{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \mathbb{1}\{T_i < \hat{f}_a(X_i) \leqslant C_i\}}{\sum_{i \in \mathcal{I}_2} \hat{w}_a(X_i) \mathbb{1}\{\hat{f}_a(X_i) \leqslant C_i\}}.$$

6: Compute the threshold $\hat{a} = \sup\{a \in \mathcal{A} \colon \sup_{a' \leqslant a, a' \in \mathcal{A}} \hat{\alpha}(a') \leqslant \alpha\}$.

Return: The calibrated LPB: $\hat{L}(\cdot) = \hat{f}_{\hat{a}}(\cdot)$.

The computational cost of our proposed procedure can be decomposed into the cost of model fitting on $\mathcal{I}_1$ and that of finding $\hat{a}$ on $\mathcal{I}_2$. The cost of the first stage heavily depends on the type of models chosen by the user. For the second stage, we first need to find the set of knots $\mathcal{A}$ defined in (5). For each $i \in \mathcal{I}_2$, finding $\bar{a}_i$ (respectively $\tilde{a}_i$) requires finding the supremum over a such that $\hat{f}_a(X_i) \leqslant \tilde{T}_i$ (respectively $\hat{f}_a(X_i) \leqslant C_i$). Since $\hat{f}_a(X_i)$ is nondecreasing in a, finding $\bar{a}_i$ or $\tilde{a}_i$ can be done efficiently via binary search. More specifically, given a tolerance level ϵ, we can obtain an ϵ-accurate solution within $O\{\log(1/\epsilon)\}$ runs. Repeating the

above for all $i \in \mathcal{I}_2$ requires $O\{|\mathcal{I}_2| \log(1/\epsilon)\}$ runs. Finally, evaluating $\hat{\alpha}(a)$ for $a \in \mathcal{A}$ and finding $\hat{a}$ requires $O(|\mathcal{I}_2|)$ runs. Overall, the computational complexity of the second stage is $O[|\mathcal{I}_2|\{1 + \log(1/\epsilon)\}]$.

3.3. *Theoretical guarantee: a double robustness result*

In this section, we establish the theoretical guarantees for the LPBs produced by Algorithm 1. In particular, we show that the LPBs enjoy a double-robustness property in the following sense: the LPBs are approximately marginally calibrated if either the censoring mechanism or the conditional quantile of survival times can be estimated well; when the latter is true, the LPBs are furthermore approximately conditionally calibrated.

Given the class of functions $\{\hat{f}_a(\cdot)\}_{a\in[0,1]}$, we define the oracle weights $w_a(x) = [\mathbb{P}\{C \geqslant \hat{f}_a(X) \mid \mathcal{I}_1,\ X = x\}]^{-1}$; here we condition on $\hat{f}_a$, i.e., the function is treated as fixed, and the following oracle quantity for any $\beta \in [0, 1]$:

$$a(\beta) = \sup\{a \in [0, 1] \colon \mathbb{P}\{T < \hat{f}_a(X) \mid \mathcal{I}_1\} \leqslant \beta\}.$$

The following two theorems develop the coverage guarantee for the LPBs.

THEOREM 3. *Fix any $\delta, \alpha \in (0, 1)$. Assume that $\hat{f}_a(x)$ is continuous in a and that, for any $a \in [0, 1]$, there exist some constant $\hat{\gamma}_a > 0$ such that $\hat{w}_a(x) \leqslant \hat{\gamma}_a$ for P_X-almost all x. Then, with probability at least $1 - \delta$ over the draw of $\mathcal{D}$, the LPB produced by Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{P}_{(X,T)\sim P}\{T \geqslant \hat{L}(X) \mid \mathcal{D}\} \\ \geqslant 1 - \alpha - \sup_{a\in[0,1]} \left(\mathbb{E}\left[\left| \frac{\hat{w}_a(X)}{w_a(X)\hat{\pi}_a} - 1 \right| \,\middle|\, \mathcal{I}_1 \right] \right. \\ \left. + \left\{ \frac{1 + \hat{\gamma}_a^2/\hat{\pi}_a^2 + \max(1, \hat{\gamma}_a/\hat{\pi}_a - 1)^2}{|\mathcal{I}_2|} \log\left(\frac{1}{\delta}\right) \right\}^{1/2} \right),\end{aligned}$$

where the probability is taken with respect to a new data point $(X, T) \sim P_{(X,T)}$, and where we define $\hat{\pi}_a = \mathbb{E}_{X\sim P_X}[\hat{w}_a(X)/w_a(X) \mid \mathcal{I}_1]$ for any $a \in [0, 1]$.

In other words, if the estimates $\hat{w}_a$ are accurate approximations of w_a, up to rescaling by a constant, then we have $\hat{w}_a(X)/w_a(X)\hat{\pi}_a \approx 1$, and approximate coverage is guaranteed. The proof of Theorem 3 is deferred to the Supplementary Material.

Next, we show that we also achieve approximate coverage when $T \mid X$ can be accurately modelled.

THEOREM 4. *Fix any $\delta, \alpha \in (0, 1)$. Assume that the same conditions as in Theorem 3 hold, and assume further that the conditional distribution of $T \mid X$ is continuous, with its conditional density upper bounded by a constant $B > 0$, and that there exists a constant $r > 0$ such that*

(a) $\sup_{\xi\in[a(\alpha),a(\alpha+r)+\psi]} w_\xi(x) \leqslant \gamma$ *and* $\sup_{\xi\in[a(\alpha),a(\alpha+r)+\psi]} \hat{w}_\xi(x) \leqslant \hat{\gamma}$ *for some constants $\psi, \gamma, \hat{\gamma} > 0$;*

(b) $\sup_{\beta\in[\alpha,\alpha+r]} \sup_{x\in\mathcal{X}}\{\max(B, 1)|\hat{f}_{a(\beta)}(x) - q_\beta(x)|\} + \hat{\gamma}\gamma\{\log(1/\delta)/|\mathcal{I}_2|\}^{1/2} \leqslant r$, *where $q_\beta(x)$ is the β-quantile of T conditional on $X = x$.*

Then, with probability at least $1-\delta$ *over the draw of* $\mathcal{D}$*, the LPB produced by Algorithm 1 satisfies, for* P_X*-almost all* x*,*

$$\mathbb{P}_{(X,T)\sim P}\{T \geqslant \hat{L}(x) \mid \mathcal{D},\ X = x\}$$
$$\geqslant 1-\alpha-\sup_{\beta\in[\alpha,\alpha+r]}\sup_{x\in\mathcal{X}}\{2B|\hat{f}_{a(\beta)}(x)-q_\beta(x)|\}-\hat{\gamma}\gamma\left\{\frac{1}{|\mathcal{I}_2|}\log\left(\frac{1}{\delta}\right)\right\}^{1/2}.$$

The proof of Theorem 4 is deferred to the Supplementary Material, where we in fact prove a more general version. The implication of Theorem 4 is that if $T \mid X$ can be modelled well, the conditional miscoverage rate will be small, which also implies that the marginal coverage rate will be small.

Remark 1. The assumption on the continuity $\hat{f}_a(x)$ in a and the boundedness on the estimated weights can simply be satisfied by choosing the appropriate class of functions in the training stage, i.e., the fitting procedure using $\mathcal{I}_1$. The additional assumption (a) requires the oracle weights $w_\beta(x)$ to be bounded at least in a neighbourhood of $a(\alpha)$; (b) is satisfied when $T \mid X$ is estimated uniformly well in a neighbourhood of α and when $|\mathcal{I}_2|$ is sufficiently large.

4. Simulations

We set up six synthetic experiments and, under each setting, we generate $N = 100$ independent and identically distributed datasets. The code for reproducing all numerical results from the simulation and the real data analysis can be found at https://github.com/zhimeir/adaptive_conformal_survival_paper. Each dataset consists of the training set $\mathcal{I}_1$, calibration set $\mathcal{I}_2$ and the test set $\mathcal{I}_3$, where $|\mathcal{I}_1| = 1000$, $|\mathcal{I}_2| = 1000$ and $|\mathcal{I}_3| = 5000$. For all experiments, the target level is $1-\alpha = 90\%$. In these experiments, we implement our proposed method with two families of bounds.

(a) DFT-adaptive-T: the candidate LPB is given by $\hat{f}_a(x) = \hat{q}_a(x)$, where $\hat{q}_a(x)$ is the estimated ath conditional quantile of T given $X = x$.
(b) DFT-adaptive-CT: the candidate LPB is given by

$$\hat{f}_a(x) = \min\{\hat{q}_a(x), \hat{q}^C_{1-\log(1/|\mathcal{I}_2|)}(x)\},$$

where $\hat{q}_a(x)$ is as before and $\hat{q}^C_b(x)$ is the estimated bth conditional quantile of C given $X = x$.

We also obtain LPBs based on parametric models and other distribution-free methods.

(i) Cox: LPBs generated by the estimated Cox model that is implemented as in Therneau (2020).
(ii) RandomForest: LPBs returned by the censored quantile regression forest (Athey et al., 2019, Li & Bradic, 2020); the implementation is based on Li & Bradic (2020).
(iii) DFT-baseline: The distribution-free LPBs obtained by applying conformal quantile regression (Romano et al., 2019) to generating bounds for $\tilde{T}$.
(iv) DFT-fixed: conformalized LPB with a fixed thresholded c_0; the implemental details are as suggested by Candès et al. (2023).

Table 1. *Parameters used in the six experimental settings:* $P_X = \mathrm{Un}([0,4]^p)$, $P_{T|X} = \exp[\mathcal{N}\{\mu(X), \sigma^2(X)\}]$

setting	p	$\mu(x)$	$\sigma(x)$	$P_{C\mid X}$
1	1	$0.632x$	2	Ex(0.1)
2	1	$3 \cdot \mathbb{1}\{x > 2\} + x \cdot \mathbb{1}\{x \leqslant 2\}$	0.5	Ex(0.1)
3	1	$2 \cdot \mathbb{1}\{x > 2\} + x \cdot \mathbb{1}\{x \leqslant 2\}$	0.5	Ex$\{0.25 + (6 + x)/100\}$
4	1	$3 \cdot \mathbb{1}\{x > 2\} + 1.5x \cdot \mathbb{1}\{x \leqslant 2\}$	0.5	lognormal$\{2 + (2 - x)/50, 0.5\}$
5	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	1	Ex$(x_{10}/10 + 1/20)$
6	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	$(x_2 + 2)/4$	Ex$(x_{10}/10 + 1/20)$

For all conformalized methods, the base algorithm for fitting the conditional quantile of $T \mid X$ is the Cox model, and a Gaussian process model is fitted to approximate $C \mid X$; this is implemented by the `GauPro` `R` package (Erickson, 2021, R Development Core Team, 2024). For each dataset, we compute the following two quantities with the test set:

$$\text{Empirical coverage} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \mathbb{1}\{\hat{L}(X_i) < T_i\},$$

$$\text{Average LPB} = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \hat{L}(X_i).$$

An ideal method would have empirical coverage $\approx 1-\alpha$, and average LPB as low as possible. We demonstrate boxplots of the empirical coverage and average LPB resulting from the 100 datasets.

4.1. *Synthetic set-up*

We consider six data-generating models, where settings 1–4 concern univariate X and settings 5–6 multivariate X. For all settings, the marginal distribution of the covariates is given by $P_X = \mathrm{Un}([0,4]^p)$; conditional on X, we generate T and C via distributions $\log T \mid X \sim \mathcal{N}\{\mu(X), \sigma^2(X)\}$ and $C \mid X \sim P_{C|X}$.

In settings 1 and 2, $p = 1$ and $C \mid X \sim \mathrm{Ex}(0.1)$; the censoring mechanism is completely exogenous. In settings 3 and 4, $p = 1$ and we allow C to depend on X. In particular, setting 3 corresponds to the example shown in the Introduction. Settings 5 and 6 consider multivariate X, where $p = 10$; in setting 5 $\sigma(x) = 1$, and in setting 6 $\sigma(x)$ depends on X. Table 1 summarizes the parameters used in the six settings.

Figure 2 shows the scatterplots of the survival time T and censoring time C against the univariate covariate X in univariate experimental settings. Settings 2–4 are more challenging than setting 1, as they all have scenarios where there are roughly two subpopulations: the subpopulation with smaller values of X has relatively higher censoring time, leading to a low censoring zone, while the subpopulation with larger values of X has comparatively higher survival time, and hence a very high censoring zone. In settings 5–6, there is a similar challenge: the distribution $P_{C|X}$ depends on X_{10} and thus we have low censoring times for certain values of X and higher censoring times for others.

4.2. *Simulation results*

Figure 3 plots the empirical coverage and average LPBs of all candidate methods under the univariate settings. First, we consider the methods without distribution-free-type

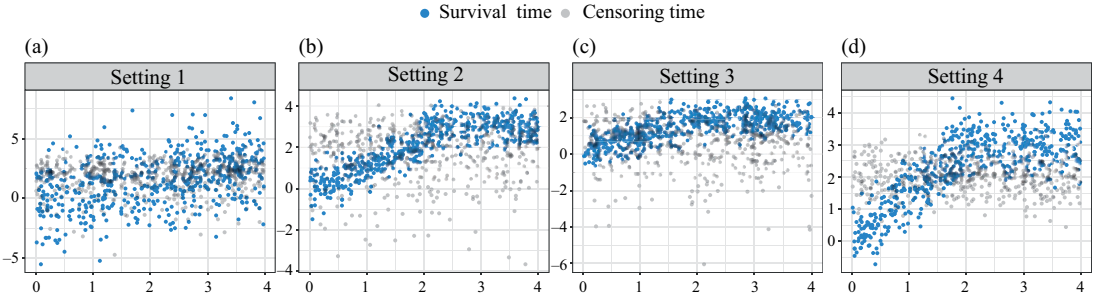

Fig. 2. Illustration of the censoring time C and the uncensored survival time T in settings 1–4 as functions of the univariate covariate X. Both C and T are on a log scale.

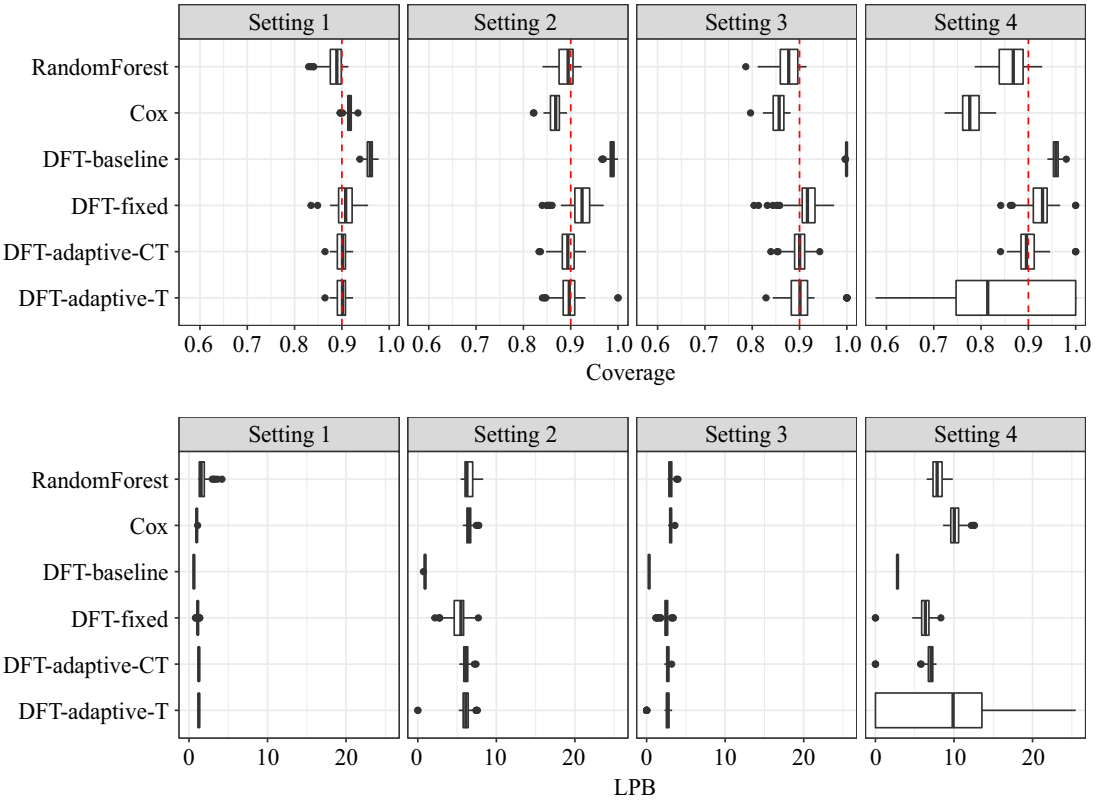

Fig. 3. Empirical coverage (top) and average LPBs (bottom) of all the candidate methods under settings 1–4, where X is univariate. The boxplot shows results from 100 independently drawn datasets. The dashed red line corresponds to the target coverage level $1 - \alpha = 90\%$.

guarantees. As we see from the figure, RandomForest exhibits undercoverage in all four settings. Even in setting 1, a simple case where the error is homogeneous and C does not depend on X, RandomForest does not return valid LPBs. Cox shows undercoverage as well, except for the simple regime of setting 1, and the miscoverage gaps are even larger than those of RandomForest in settings 2, 3 and 4.

Next we consider the DFT methods. The LPBs returned by DFT-baseline are very conservative in all four settings due to the censoring issue, as we expected, recall from our discussion in § 1.3 that this method covers T by covering the censored time $\tilde{T}$. DFT-fixed LPBs

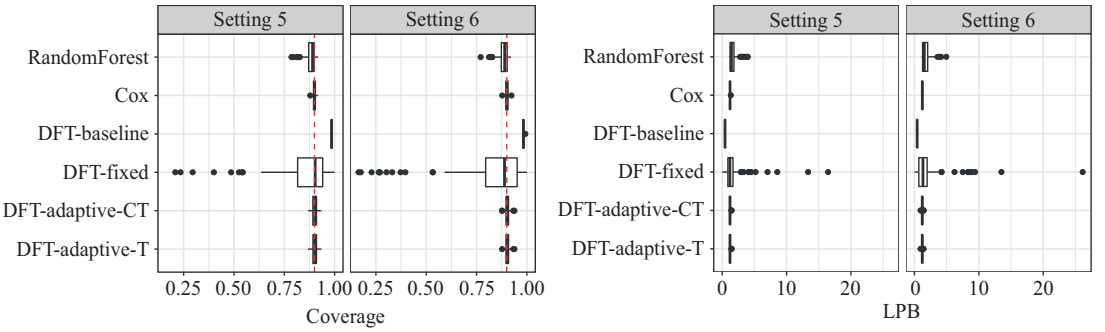

Fig. 4. Boxplots of the empirical coverage (left) and average LPBs (right) in the multivariate experimental settings. The details are otherwise the same as in Figure 3.

are more conservative than our proposed DFT-adaptive-CT LPBs, especially in settings 2–4 where the relationship between C and T changes drastically in different subpopulations. Finally, we can see that the canonical version of our method, DFT-adaptive-T, exhibits high variability in settings 3 and 4, verifying our statement on stability in § 3.2 and highlighting the potential advantages of DFT-adaptive-CT.

Next, Fig. 4 demonstrates the results under the multivariate settings. Here, we observe that RandomForest shows slight undercoverage under both settings; Cox performs well and does not undercover, but does not offer a distribution-free coverage guarantee. Turning to the DFT methods, DFT-baseline is again very conservative, and DFT-fixed LPBs exhibit high variability. Both of our proposed methods, DFT-adaptive-T and DFT-adaptive-CT, are able to achieve exact coverage, and the variability is much lower than that of DFT-fixed LPBs.

Finally, we show in Fig. 5 the running time of all the candidate methods under setting 3, the example in § 1.5, with different choices of n. We can see that DFT-fixed, DFT-adaptive-T and DFT-adaptive-CT are more computationally expensive than the other methods; the running times of DFT-fixed and DFT-adaptive-CT are comparable, while that of DFT-adaptive-T is somewhat shorter.

5. Real data application

In this section, we apply our proposed method to predicting users' active time on a mobile app with a publicly available dataset. The data are available from `https://www.kaggle.com/datasets/bhuvanchennoju/mobile-usage-time-prediction?select=pings.csv`. This dataset records the time stamps of pings for a cohort of 2476 users in a shared window of three weeks, where a ping represents a login activity or a received message. As shown in Fig. 6(a), a user's pings gathered during an active day form a line segment, whose length is proportional to the span of active time during that day; the time span is standardized so that the total time window is mapped to the interval $[0, 21]$ to represent the total number of days. The number of line segments and the length of line segments vary for different users, reflecting different types of user behaviour. For each user, the time is recorded from the user's first active day, i.e., if the first active time for a particular user is 1.5 then the active time sequence for this user is shifted by $\lfloor 1.5 \rfloor = 1$ and is censored at time $C = 21 - 1 = 20$.

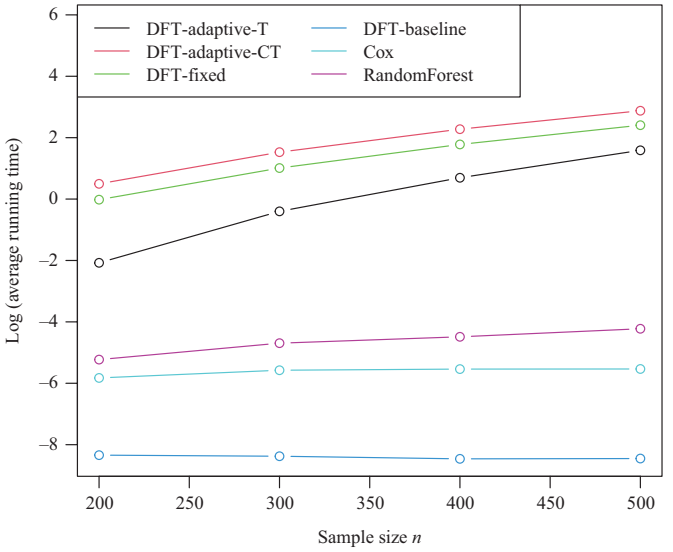

Fig. 5. Running times of all candidate methods in setting 3.

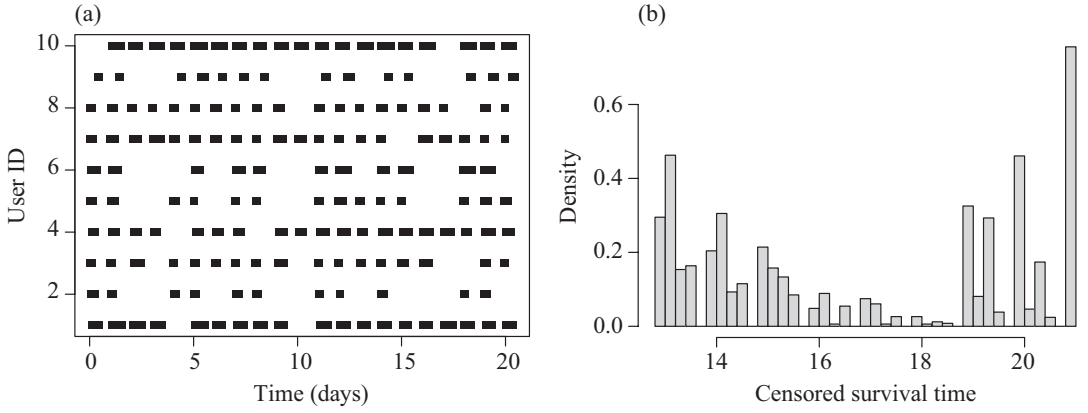

Fig. 6. (a) Active time in three weeks for $i = 1, \ldots, 10$. (b) Histogram of censored survival times for all users.

With this dataset, we focus on predicting the beginning of a user's 14th active day. In practice, the prediction lower bounds can be informative if, for instance, the mobile app wishes to launch a promotion, offering a discount for in-app purchases at the beginning of a user's 14th active day. For a user who is active for less than 14 days within the time window, the survival time is therefore censored and only $\tilde{T} = \min(C, T)$ can be observed. Figure 6(b) is the histogram of the censored survival time. Besides the time stamps, there are three covariates in this dataset related to users' characteristics: X_1 (gender), X_2 (age) and X_3 (number of children).

To implement the method, we begin by choosing $|\mathcal{I}_1| = 500$ data points as the training set, and keep this set fixed throughout. Among the remaining data points, for 50 independent random trials, we sample $|\mathcal{I}_2| = 500$ data points as the calibration set and another $|\mathcal{I}_3| = 500$ as the test set, uniformly without replacement. All the methods are applied with the target level $1-\alpha$ at 90%. Since the true survival times for censored data points are not available, we instead empirically evaluate the upper and lower bounds of the coverage rate: we compute

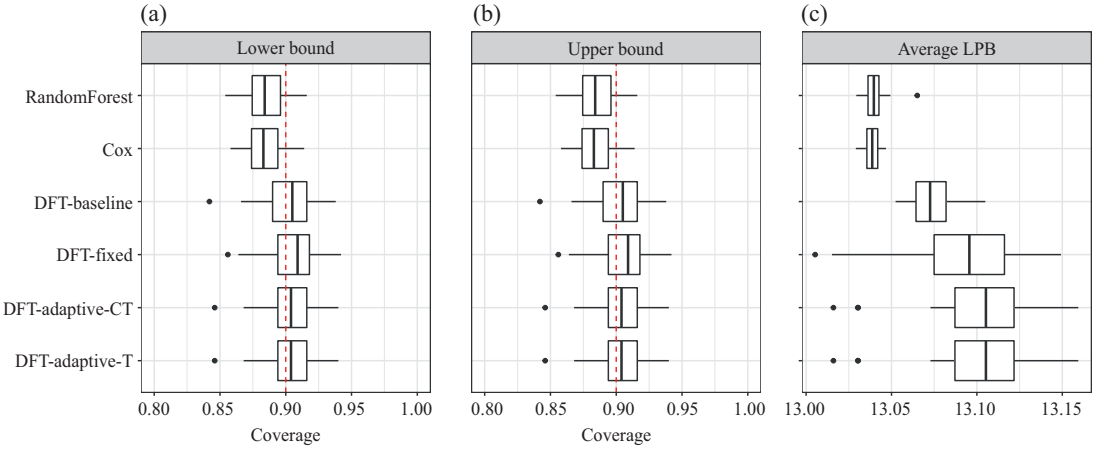

Fig. 7. (a) Lower bound β_{lo} of the empirical coverage rate. (b) Upper bound β_{hi} of the empirical coverage rate. (c) Average LPBs.

$\beta_{\text{lo}} := \mathbb{P}\{\tilde{T} \geqslant \hat{L}(X)\} \leqslant \mathbb{P}\{T \geqslant \hat{L}(X)\}$ and also $\beta_{\text{hi}} := 1 - \mathbb{P}\{\tilde{T} < \hat{L}(X),\ T \leqslant C\} \geqslant \mathbb{P}\{T \geqslant \hat{L}(X)\}$, so that, by construction, β_{lo} is an underestimate of our target coverage rate and β_{hi} is an overestimate.

The upper and lower bounds for methods in comparison are reported in Fig. 7. Since this setting has a low censoring rate (19.7%), the difference between DFT-adaptive-T and DFT-adaptive-CT is negligible. Both DFT-adaptive-T and DFT-adaptive-CT attain nearly exact coverage at 90%, while Cox and RandomForest have coverage below the target level. In comparison, although DFT-baseline has only slightly inflated coverage, the average LPB is lower than those of our methods and is thus less accurate in practice; in the meantime, DFT-fixed has a coverage rate slightly higher than 90% and shows larger variance than our methods with adaptive cut-offs.

6. Discussion

As in Candès et al. (2023), this work has primarily focused on Type-I censoring, where the censoring time for each individual is assumed observable; this is typically the case when the censoring time is the termination of a study. Another common type of censoring time is loss-to-follow-up censoring. When the event is death, the loss-to-follow-up censoring time is not observed for patients who did not survive, and our method no longer applies. As discussed by Candès et al. (2023), our method can however provide informative LPBs beyond the setting of Type-I censoring: when we have both the end-of-study censoring time C_{end} and the loss-to-follow-up censoring time C_{loss}, the censored survival time is then given by $\tilde{T} = T \wedge C_{\text{end}} \wedge C_{\text{loss}}$. Under the assumption that $(T, C_{\text{loss}}) \perp\!\!\!\perp C_{\text{end}} \mid X$, we can treat $T' := T \wedge C_{\text{loss}}$ as the true survival time and apply our procedure, producing an LPB on T'. We can thus alleviate the conservativeness caused by C_{end}, especially in studies with short duration.

The theoretical guarantees shown in this work focus on constructing the PAC-type LPB. It would also be of interest to see if one can derive marginal guarantees for the proposed method, where the weighted conformal inference technique is not applicable. Recent work by Angelopoulos et al. (2023) on a related problem suggests tools for converting a PAC-type bound to a finite-sample bound in expectation, and may be applicable to the survival analysis setting as well.

As with many double-robustness-type results, our theoretical guarantees rely on high accuracy of our estimate of either the conditional distribution of $C \mid X$ or of $T \mid X$, but it may be possible to establish a better bound where moderately accurate estimates of both distributions contribute multiplicatively to a single unifying bound; this may be more relevant to practical settings, where we might expect moderate accuracy for each estimation problem. Finally, as discussed earlier, the cut-off introduces a variance-bias trade-off; with a large cut-off, the observed survival time is closer to the true survival time, but the effect sample size is reduced, and vice versa. It would be interesting to quantitatively characterize this phenomenon, and derive an optimal choice of candidate LPBs based on this characterization.

Acknowledgement

Ren and Barber were supported by the Office of Naval Research (N00014-20-1-2337). Barber was also supported by the National Science Foundation (DMS-1654076 and DMS-2023109).

Supplementary material

The Supplementary Material contains proofs of Theorems 3 and 4.

References

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I. & Lei, L. (2022). Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv:* 2110.01052v5.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L. & Schuster, T. (2023). Conformal risk control. *arXiv:* 2208.02814v3.

Athey, S., Tibshirani, J. & Wager, S. (2019). Generalized random forests. *Ann. Statist.* **47**, 1148–78.

Bates, S., Angelopoulos, A., Lei, L., Malik, J. & Jordan, M. (2021). Distribution-free, risk-controlling prediction sets. *J. ACM* **68**, 1–34.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.* **43**, 45–57.

Candès, E. J., Lei, L. & Ren, Z. (2023). Conformalized survival analysis. *arXiv:* 2103.09763v3.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* B **34**, 187–202.

Erickson, C. (2021). `GauPro`: Gaussian process fitting. `R` package version 0.2.4.

Faraggi, D. & Simon, R. (1995). A neural network model for survival data. *Statist. Med.* **14**, 73–82.

Fleming, T. R. & Lin, D. (2000). Survival analysis in clinical trials: past developments and future directions. *Biometrics* **56**, 971–83.

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**, 3001–8.

Harrell, F. E., Jr. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer.

Jin, Y., Ren, Z. & Candès, E. J. (2022). Sensitivity analysis of individual treatment effects: a robust conformal inference approach. *arXiv:* 2111.12161v2.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data* . Hoboken, NJ: John Wiley and Sons.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. & Kluger, Y. (2016). Deep survival: a deep Cox proportional hazards network. *BMC Med. Res. Method* **18**, 1–12.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Asymptotic Statistics*, Ed. P. Mandl and M. Hušková, pp. 349–59. Heidelberg: Springer.

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. *Sci. Rep.* **7**, 1–8.

Lei, L. & Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *arXiv:* 2006.06138v2.
Leung, K.-M., Elashoff, R. M. & Afifi, A. A. (1997). Censoring issues in survival analysis. *Ann. Rev. Public Health* **18**, 83–104.
Li, A. H. & Bradic, J. (2020). Censored quantile regression forest. In *Proc. 23rd Int. Conf. Artif. Intel. Statist.*, pp. 2109–19. PMLR.
Muenchow, G. (1986). Ecological use of failure time analysis. *Ecology* **67**, 246–50.
Murphy, S., Rossini, A. & van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Am. Statist. Assoc.* **92**, 968–76.
R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Romano, Y., Patterson, E. & Candes, E. (2019). Conformalized quantile regression. In *Proc. 33rd Int. Conf. Neural Info. Proces. Syst.*, pp. 3543–53. Red Hook, NY: Curran Associates.
Singh, R. & Mukhopadhyay, K. (2011). Survival analysis in clinical trials: basics and must know areas. *Perspect. Clin. Res.* **2**, 145.
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. *Statist. Med.* **16**, 385–95.
Tibshirani, R. J., Foygel Barber, R., Candes, E. & Ramdas, A. (2019). Conformal prediction under covariate shift. In *Proc. 33rd Int. Conf. Neural Info. Proces. Syst.*, pp. 2530–40. Red Hook, NY: Curran Associates.
Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proc. Asian Conf. Mach. Learn.*, pp. 475–90. PMLR.
Vovk, V., Gammerman, A. & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Boston, MA: Springer.
Wang, P., Li, Y. & Reddy, C. K. (2019). Machine learning for survival analysis: a survey. *ACM Comp. Surveys* **51**, 1–36.
Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statist. Med.* **11**, 1871–9.

[*Received on* 12 *November* 2022. *Editorial decision on* 16 *August* 2023]