T-SCI: A Two-Stage Conformal Inference Algorithm with Guaranteed Coverage for Cox-MLP

Jiaye Teng [*1] **Zeren Tan** [*1] **Yang Yuan** [1 2]

Abstract

It is challenging to deal with censored data, where we only have access to the incomplete information of survival time instead of its exact value. Fortunately, under linear predictor assumption, people can obtain guaranteed coverage for the confidence band of survival time using methods like Cox Regression. However, when relaxing the linear assumption with neural networks (e.g., Cox-MLP (Katzman et al., 2018; Kvamme et al., 2019)), we lose the guaranteed coverage. To recover the guaranteed coverage without linear assumption, we propose two algorithms based on conformal inference under strong ignorability assumption. In the first algorithm *WCCI*, we revisit weighted conformal inference and introduce a new non-conformity score based on partial likelihood. We then propose a two-stage algorithm *T-SCI*, where we run WCCI in the first stage and apply quantile conformal inference to calibrate the results in the second stage. Theoretical analysis shows that T-SCI returns guaranteed coverage under milder assumptions than WCCI. We conduct extensive experiments on synthetic data and real data using different methods, which validate our analysis.

1. Introduction

In survival analysis, censoring indicates that the value of interest (survival time) is only partially known (e.g., the information can be $t > 5$ instead of $t = 7$). It is common and inevitable in numerous fields, including medical care (Robins & Finkelstein, 2000; Klein & Moeschberger, 2006), astronomy (Feldmann, 2019), finance (Bellotti & Crook, 2009), etc. It is an annoying issue since ignoring or deleting censored data causes bias and inefficiency (Nakagawa & Freckleton, 2008), as illustrated in Figure 1.

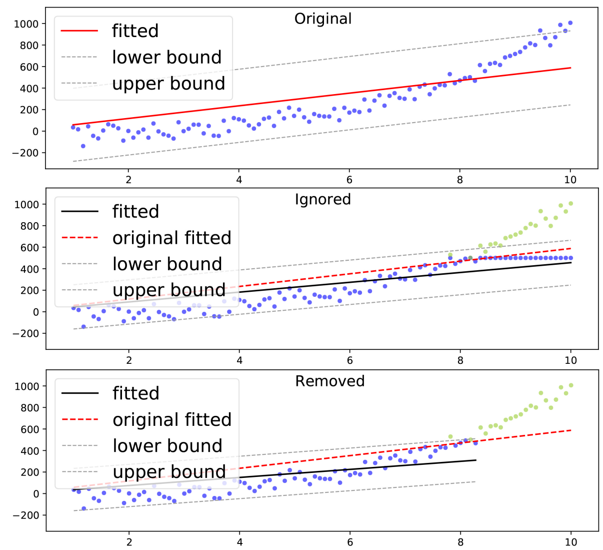


Figure 1. **Bias in censoring.** Ignoring or deleting censored data (light green points) causes bias compared to the ground truth linear approximation (red line), where the linear approximation under censoring (black line) does not overlap the ground truth (red line).

When dealing with censored data, we usually focus on the *confidence band* of the survival time since confidence bands give a more conservative estimation than point estimation. Under linear assumptions on covariate effect (See Assumption 1), one can derive the survival time distribution using Cox regression by asymptotic normality of linear coefficient. It further leads to guaranteed coverage, meaning that survival time provably falls into the confidence band with high probability (larger than the given confidence level).

However, the linear assumption broadly harms its performances and restricts its applications. After all, the reality is not always entirely linear. To relax the linear assumption, Katzman et al. (2018); Kvamme et al. (2019) applies neural networks into Cox regression (Cox-MLP), yielding the best performance in terms of some metrics such as Brier score and binomial log-likelihood. Unfortunately, the confidence band in Cox-MLP has no guaranteed coverage since one cannot expect neural network *converges* to the expected

[*]Equal contribution [1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2]Shanghai Qi Zhi Institute, Shanghai, China. Correspondence to: Yang Yuan <yuanyang@tsinghua.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

function, not to mention the asymptotic normality.

In general, we can use *conformal inference* to recover the confidence band with guaranteed coverage by splitting the dataset into training and calibration set (Vovk et al., 2005; Nouretdinov et al., 2011; Lei & Candès, 2020). One of its advantages is that conformal inference does not harm the model performance since it is post-hoc. Therefore, we propose to apply conformal inference into Cox-MLP.

When applying conformal inference into Cox-MLP, there are several problems to be solved. Firstly, Cox regression does not return survival time explicitly, which requires a modification of the non-conformity score. Secondly, censoring causes *covariate shift* under strong ignorability, meaning that the covariate distribution differs in censored and uncensored data. Therefore, we cannot apply conformal inference directly. Thirdly, we need to consider the estimation error and provide theoretical guarantees for the coverage.

In this paper, we propose a new non-conformity score under strong ignorability assumption (which is a standard assumption in weighted conformal inference. We refer to more details in Section 3) based on the partial likelihood of Cox regression. This non-conformity score does not need an explicit estimation of the survival time. We then apply weighted conformal censoring inference (WCCI), a weighted conformal inference based on this non-conformity score inspired by Tibshirani & Foygel (2019) to deal with the covariate shift problem. Furthermore, inspired by Romano et al. (2019), we provide a two-stage conformal inference (T-SCI) which returns "nearly perfect" coverage, meaning that the coverage has not only guaranteed lower bound but also upper bound. Inspired by (Lei & Candès, 2020), we provide theoretical guarantees for both WCCI and T-SCI algorithms.

We summarize our contributions as follows:

- We provide coverage for Cox-MLP in WCCI based on weighted conformal inference frameworks by introducing a new non-conformity score.
- We further propose a T-SCI algorithm based on the quantile conformal inference framework. We show that T-SCI returns nearly perfect guaranteed coverage, namely, theoretical guarantees for coverage's lower and upper bound.
- We conduct extensive experiments on both synthetic data and real-world data, showing that the T-SCI-based algorithm outperforms other approaches in terms of empirical coverage and interval length.

2. Related work

Censored data analysis. An early analysis of censored data can be dating back to the famous Kaplan–Meier estimator Kaplan & Meier (1958). However, this approach is valid only when all patients have the same survival function. Therefore, several individual-level analysis is proposed, such as proportional hazard model (Breslow, 1975), accelerated failure time model (Wei, 1992) and Tree-based models (Zhu & Kosorok, 2012; Li & Bradic, 2020).

On the other hand, researchers apply machine learning techniques to deal with censored data (Wang et al., 2019). For example, random survival forests (Ishwaran et al., 2008) train random forests using the log-rank test as the splitting criterion. Moreover, DeepHit (Lee et al., 2018) apply neural networks to estimate the probability mass function and introduce a ranking loss. This paper mainly focuses on the *Cox-based model*, one of the famous proportional hazard model branches.

Cox regression was first proposed in Cox (1972), which is a semi-parametric method focusing on estimating the hazard function. Among all its extensions, Akritas et al. (1995) first proposed using a one-hidden layer perceptron to replace the linear predictor of the coefficient. However, it generally failed mainly due to the low expressivity of the one-hidden layer perceptron (Xiang et al., 2000; Sargent, 2001). Therefore, Katzman et al. (2018) proposed to use the multi-layer perceptron instead of the one-layer perceptron (DeepSurv). Furthermore, (Kvamme et al., 2019) generalize the idea to the non-proportional hazard settings. In this paper, we unify their names as *Cox-MLP* when the context is clear. However, this line of work lacks a theoretical guarantee. This paper tries to fill this blank and propose the first guaranteed coverage of the survival time.

Conformal inference was pioneered by Vladimir Vovk and his collaborators [e.g., Vovk et al. (2005); Shafer & Vovk (2008); Nouretdinov et al. (2011)], focusing on the inference of response variables by splitting a training fold and a calibration fold. There are several variations of conformal inference. For example, weighted conformal inference (Tibshirani & Foygel, 2019) focus on dealing with the covariate shift phenomenon, and quantile conformal inference (Romano et al., 2019) returns coverage with not only an upper bound but also the lower bound guaranteed coverage. Conformal inference, as well as its variations, are widely studied and used in Lei et al. (2013); Lei & Wasserman (2014); Lei et al. (2018); Barber et al. (2019b); Sadinle et al. (2019); Romano et al. (2020); Angelopoulos et al. (2020).

Recently, Lei & Candès (2020) apply conformal inference under counterfactual settings and derive a double robust guarantee for their proposed methods. Our work is partially inspired by Lei & Candès (2020) but considers a different censoring setting. Furthermore, we propose a different algorithm and derive a "nearly perfect" guaranteed coverage. A very recent work (Candès et al., 2021) focuses on a similar censoring scenario. Unlike our approach (T-SCI), Candès et al. (2021) relax the strong ignorability assumption but require all information of censoring time to obtain confidence bands. We emphasize that it is still an open problem on deriving guaranteed confidence bands under general censoring scenarios.

There are some approaches which apply conformal inference under censoring settings. For example, Bostr et al. (2017); Boström et al. (2019) apply conformal inference into random survival forests, and Chen (2020) derive the confidence band for DeepHit. However, one cannot apply these approaches to Cox-based models since Cox-based models do not explicitly return a predicted survival time.

3. Preliminary

Denote $X \sim \mathcal{P}_X \subset \mathbb{R}^d$ the covariate, $T \sim \mathcal{P}_T \subset \mathbb{R}$ the survival time, and $C \sim \mathcal{P}_C \subset \mathbb{R}$ the censoring time. Denote the joint distribution by $\mathcal{P}_{XTC}$, its marginal distribution by $\mathcal{P}_{TX}$, and its conditional distribution by $\mathcal{P}_{T|X}$. The survival time cannot be observed when it is larger than the censoring time. Therefore, observed time is the minimum of censoring time and survival time. Let $Y \in \mathbb{R}$ be the observed time with uncensoring indicator $\Delta \in \mathbb{R}$, then

$$Y = \min\{T, C\},\ \Delta = \mathbb{I}_{\{T \le C\}}.$$

Denote the dataset by $\mathcal{Z} = \{X_i, Y_i, \Delta_i\}_{i \in \mathcal{I}}$ where we can only access the covariate, the observed time, and the censoring indicator. However, the value of interest is the survival time T. Therefore, the censored data is incomplete due to the information loss when $\Delta = 0$. We next introduce how censoring happens, namely, the censoring mechanism.

Censoring Mechanism. In this paper, we consider the censoring regimes with strong ignorability assumption, namely $T \perp \Delta \mid X$. We will further discuss the strong ignorability assumption in Appendix C.

However, note that there can be covariate shift under such censoring regimes, namely, the distributions of covariate X under censoring and non-censoring are different

$$(X \mid \Delta = 1) \overset{d}{\neq} (X \mid \Delta = 0)\,.$$

We refer to Figure 2 for an illustration. Usually, we estimate the distribution of the survival time via the dataset $\mathcal{D}$ by *Cox regression*. We denote $F_T(t)$ the survival time's CDF.

Cox Regression. In Cox regression, we focus on two important terms *survival function* $S_T(t)$ and *cumulative hazard function* $\Lambda_T(t)$, as defined in Equation 1. We emphasize that they are defined with respect to the survival time T instead of the observed time Y.

$$S_T(t) \triangleq 1 - F_T(t),\ \Lambda_T(t) \triangleq -\log S_T(t). \tag{1}$$

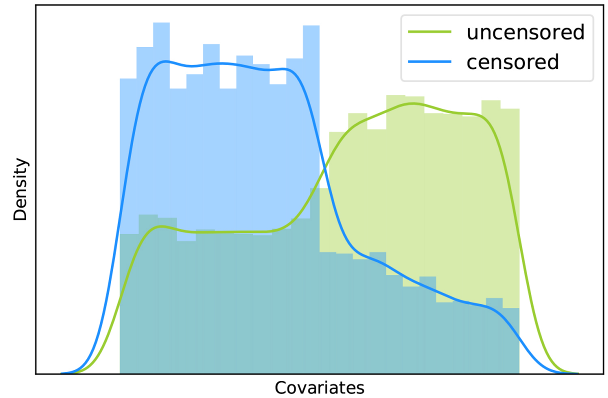


Figure 2. **Covariate shift illustration.** We show distributions of x_1 with (blue) and without (green) censoring under simulation data. Obviously, censored and uncensored data have different distributions, i.e. covariate shift[1].

When the context is clear, we omit the subscript T and denote the above function by $F(t)$, $S(t)$, and $\Lambda(t)$, respectively. Cox-based models usually require *proportional hazard* assumption, formally stated in Assumption 1.

Assumption 1 (Proportional Hazard) *For each individual i, we assume*

$$\Lambda_i(t; X_i) = \Lambda_0(t)\exp\left(g(X_i)\right) \tag{2}$$

where $\Lambda_0(t)$ is the baseline cumulative hazard function, and $g(X_i)$ is the individual effect named as predictor.

Remark: There are several non-proportional hazard Cox models which replace $g(x_i)$ with $g(x_i, t)$ (e.g., Kvamme et al. (2019)). Although our proposed algorithm can be directly generalized to non-proportional settings, we only consider proportional hazard models for clarity in this paper.

Specifically, Cox regression solves the case when the predictor $g(\cdot)$ is linear by maximizing *partial log-likelihood* $l(g)$, defined in Equation 3.

$$l(g) \triangleq \sum_{j:\Delta_j = 1} \log\left(\sum_{k \in R(T_j)} \exp[g(X_k) - g(X_j)]\right), \tag{3}$$

where $R(T_j)$ is the set of all individuals at risk at time T_j- (the observed time is no less than T_j), and $g(X_j) = X_j^\top \beta$ is the linear predictor.

For the case when $g(\cdot)$ is not linear, Lee et al. (2018) and Kvamme et al. (2019) propose Cox-MLP which uses neural networks to replace the linear predictor $g(X_i)$. Concretely, they use the negative partial log-likelihood as the training loss, with a penalty on the complexity of $g(\cdot)$.

[1]The simulation dataset is from Kvamme et al. (2019)

Conformal inference. Conformal inference does post-hoc estimation based on the splitting of training and calibration fold. One trains a model μ using the training fold and then calculates the *non-conformity score* V_i on the calibration fold. A commonly used non-conformity score is the absolute error $V_i = |T_i - \mu(X_i)|$ where (X_i, T_i) is the i_{th} calibration sample. When designing the non-conformity score, one of the most important characteristics is *exchangeability*.

Assumption 2 (Exchangeability) *For $n \geq 1$ random variables $V_1, \ldots, V_n$, they satisfy exchangeability if*

$$(V_1, \ldots, V_n) \stackrel{d}{=} (V_{\pi(1)}, \ldots, V_{\pi(n)})$$

for any permutation $\pi : [n] \to [n]$, where $\stackrel{d}{=}$ means they have the same distribution, and $[n] = \{1, 2, \ldots, n\}$.

Exchangeability is weaker than independence since independence implies exchangeability. Under the exchangeability assumption on the non-conformity score, we reach a theoretical guarantee of the confidence band. In this paper, we mainly focus on two varieties of conformal inference, *weighted conformal inference* (to deal with covariate shift, Lemma 1) and *quantile conformal inference* (to return a nearly perfectly guarantee, Lemma 2), respectively.

Lemma 1 (WCI, Tibshirani & Foygel (2019) Theorem 2) *Assume that the data $(X_i, T_i), i \in [n+1]$ are weighted exchangeable with weight $w_1, \ldots, w_{n+1}$, then the returned confidence band $\widehat{C}_n$ has guaranteed coverage:*

$$\mathbb{P}(T_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

Lemma 2 (QCI, Romano et al. (2019) Theorem 1) *Assume that the data $(X_i, T_i), i \in [n+1]$ are exchangeable, and the non-conformity scores are almost surely distinct, then the returned confidence band $\widehat{C}_n$ is nearly perfectly calibrated:*

$$1 - \alpha \leq \mathbb{P}(T_{n+1} \in \widehat{C}_n(X_{n+1})) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_{ca}| + 1},$$

where $\mathcal{I}_{ca}$ denotes the number of calibration samples.

Remark. One may wonder why not use $S(t)$ to return the confidence band in Cox-MLP. People usually do so in practice, but the confidence band has no theoretical guarantee in Cox-MLP. To derive the theoretical guarantee under $S(t)$, one needs to show the convergence of the predictor $g(X_i)$. However, it cannot be proved unless the generalization guarantee of neural networks is obtained.

4. Confidence band for the Survival Time

In this section, we derive the confidence band for the survival time. We start by analyzing the basic properties and the critical ideas before proposing the algorithm.

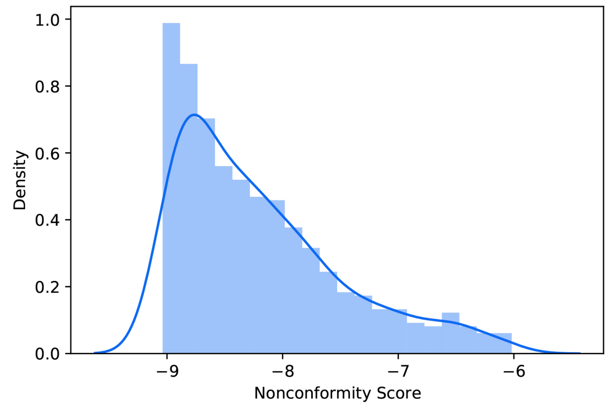


Figure 3. **Non-conformity score distribution.** We show the distribution of the non-conformity score[2], which has a well-shaped single peak (more stable, see Appendix C for more details). The dark blue line is the KDE approximation.

The non-conformity score. In traditional conformal inference, the most commonly used non-conformity score is the absolute form $V_i = |T_i - \hat{T}_i|$. That is because we can directly derive the confidence band of T_i based on the band of the non-conformity score V_i. Unfortunately, it is hard to compute $\hat{T}_i$ in Cox-based models since they output the survival hazard function instead of the survival time, which requires a modification of the non-conformity score.

Inspired by the standard Cox regression, we introduce a non-conformity score based on partial likelihood. Specifically, we use the sample partial log-likelihood as the non-conformity score (shown in Equation 4). Compared to the absolute non-conformity score, the newly proposed non-conformity score $V_i = V_g(X_i, T_i)$ do not need to calculate $\hat{T}_i$ explicitly. Figure 3 illustrates the distribution of the non-conformity score using a simulation dataset.

$$V_i = \log\left(\sum_{k \in R(T_i)} \exp[g(X_k) - g(X_i)]\right). \tag{4}$$

The incomplete data. Compared to the i.i.d. (independent and identically distributed) data, censored data is incomplete. Notice that we can only calculate the non-conformity scores of the uncensored data in Equation 4 since we cannot obtain the exact value of T_i of censored data. As a result, the returned confidence band is guaranteed under uncensored distribution $\mathcal{P}_{X|\Delta=1}$. However, as mentioned before, there might be a distribution shift between censored data and uncensored data, leading to the requirement of *weighted conformal inference* (Tibshirani & Foygel, 2019).

In weighted conformal inference, one needs to calculate the

[2]The simulation dataset is from Kvamme et al. (2019)

Algorithm 1 WCCI: Weighted Conformal Censoring Inference

Input: Level α
Input: data $\mathcal{Z} = (X_i, Y_i, \Delta_i)_{i\in\mathcal{I}}$, testing point X'
Input: function $\hat{w}(x;\mathcal{D})$ to fit the weight function at x using $\mathcal{D}$ as data

1. Randomly split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{tr} \triangleq (X_i, Y_i, \Delta_i)_{i\in\mathcal{I}_{tr}}$ and a calibration fold $\mathcal{Z}_{ca} \triangleq (X_i, Y_i, \Delta_i)_{i\in\mathcal{I}_{ca}}$
2. Use $\mathcal{Z}_{tr}$ to train $\hat{g}(\cdot)$ to estimate the predictor function
3. For each $i \in \mathcal{I}_{ca}$ with $\Delta_i = 1$, compute the non-conformity score $V_i = \log(\sum_{k\in R(T_i)\cap\mathcal{I}_{tr}} \exp[\hat{g}(X_k) - \hat{g}(X_i)])$
4. For each $i \in \mathcal{I}_{ca}$ with $\Delta_i = 1$, compute the weight $W_i = \hat{w}(X_i;\mathcal{Z}_{tr})$
5. Calculate the normalized weights $\hat{p}_i = \frac{W_i}{\sum_{i\in\mathcal{I}_{ca}} W_i + \hat{w}(X';\mathcal{Z}_{tr})}$ and $\hat{p}_\infty = \frac{\hat{w}(X';\mathcal{Z}_{tr})}{\sum_{i\in\mathcal{I}_{ca}} W_i + \hat{w}(X';\mathcal{Z}_{tr})}$
6. Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\sum_{i\in\mathcal{I}_{ca}} \hat{p}_i\delta_{V_i} + \hat{p}_\infty\delta_\infty$
7. Calculate $T^u(X')$ as the smallest value such that its conformity score V' (dependent on $T^u(X')$) is larger than $Q_{1-\alpha}$

Output: $\hat{C}^1(X') = [0, T^u(X')]$.

weight w based on $w(x)$, where

$$w(x) = \frac{\mathrm{d}\mathcal{P}_{\mathrm{X}}(\mathrm{x})}{\mathrm{d}\mathcal{P}_{\mathrm{X}|\Delta=1}(\mathrm{x})} = \frac{\mathbb{P}(T=1)}{\mathbb{P}(T=1 \mid X=x)}. \quad (5)$$

Intuitively, $w(x)$ helps transfer the guarantee over $\mathcal{P}_{X|\Delta=1}$ to $\mathcal{P}_X$. In the following of the paper, we use the regularized weight p_i in the algorithm, namely

$$p_i = \frac{w(X_i)}{\sum_{i\in\mathcal{I}_{ca}} w(X_i) + w(X')},$$

where we denote X_i as the samples from the calibration fold and X' as the testing point.

Exchangeablility. In conformal inference, we assume that the non-conformity score satisfies exchangeablility (See Assumption 2). However, as shown in Equation 4, there is a summation term in the non-conformity score, where we need to sum up all the samples at risk at time T_j-. Unfortunately, it breaks the exchangeablility when we use the at-risk samples in the calibration fold. As an alternative, we use the at-risk samples in the training fold. The non-conformity scores in the calibration fold $V_i,\ i \in [n]$ then satisfies exchangeablility given the training fold (See Equation 6).

$$(V_1, \ldots, V_n \mid \mathcal{D}_{tr}) \stackrel{d}{=} (V_{\pi(1)}, \ldots, V_{\pi(n)} \mid \mathcal{D}_{tr}), \quad (6)$$

with arbitrary permutation π.

The reconstruction of confidence band. Based on the non-conformity score proposed in Equation 4, we can reconstruct the confidence band for the survival time. We remark that we calculate the one-sided band $[0, T^u(X')]$ although the two-sided band directly follows. For a new sample X', we first calculate the $(1-\alpha)$ quantile of (weighted) distribution of the non-conformity score, denoted as $Q_{1-\alpha}$. We then calculate the smallest $T^u(X')$ that makes its non-conformity score $V' = V_{\hat{g}}(X', T)$ larger than $Q_{1-\alpha}$, formally,

$$Q_{1-\alpha} = \text{Quantile}\left(1-\alpha, \sum_{i\in\mathcal{I}_{ca}} p_i V_i\right)$$
$$T^u(X') = \inf\left\{T : V_{\hat{g}}(X', T) \geq Q_{1-\alpha}\right\}.$$

Algorithm. Based on the discussions above, we conclude our algorithm in Algorithm 1. In training process, we use training fold to train the estimated predictor $\hat{g}(X)$ (defined in Equation 2) and calculate the weight function (defined in Equation 5). We then use calibration fold to calculate the non-conformity score. We finally construct confidence band $[0, T^u(X')]$ for a given testing point X'.

Robustness on $\hat{w}$. The weighted conformal inference has guaranteed coverage under the true weight $w(x)$ as shown in Lemma 1. However, we can only obtain its estimator $\hat{w}(x)$ in practice. In the following Theorem 4.1, we prove how the estimation influences the coverage. Note that as $\lim_{|\mathcal{Z}_{tr}|\to\infty} |\hat{w}(x) - w(x)| \to 0$, the coverage can be provably larger than $1-\alpha$.

Theorem 4.1 (Provable Guarantee) *Let $\hat{w}(x)$ be an estimate of the weight $w(x)$. Assume that $\mathbb{E}[\hat{w}(X)|\mathcal{Z}_{tr}] = 1$ and $\mathbb{E}[w(X)] = 1$. Denote $\hat{C}_n^1(x)$ as the output band of Algorithm 1 with n calibration samples, then for a new data X', its corresponding survival time satisfies*

$$\lim_{n\to\infty} \mathbb{P}\left(T' \in \hat{C}_n^1(X')\right) \geq 1 - \alpha - \frac{1}{2}\mathbb{E}|\hat{w}(X) - w(X)|,$$

where the probability $\mathbb{P}$ on the left hand side is taken over $(X', T') \sim \mathcal{P}_X \times \mathcal{P}_{T|X}$, and all the expectation operators $\mathbb{E}$ are taken over $X \sim \mathcal{P}_{X|\Delta=1}$.

Theorem 4.1 proves the lower bound for WCCI. However, it is insufficient to derive the upper bound under the weighted conformal inference framework. To derive the upper bound, we propose the algorithm T-SCI in the next section.

Algorithm 2 T-SCI: Two-Stage Conformal Inference

Input: Level α
Input: additional data $\mathcal{Z}_{ca2} = (X_i, Y_i, \Delta_i)_{i \in \mathcal{I}_{ca2}}$, testing point X'
Input: First-stage band $[\hat{q}_{\alpha_{lo}}(X'; \mathcal{Z}_{tr}, \mathcal{Z}_{ca1}), \hat{q}_{\alpha_{hi}}(X'; \mathcal{Z}_{tr}, \mathcal{Z}_{ca1})]$ output from Algorithm 1 [3]
Input: function $\hat{w}(x; \mathcal{D})$ to fit the weight function at x using $\mathcal{D}$ as data
1: For each $i \in \mathcal{I}_{ca2}$ with $\Delta_i = 1$, compute the non-conformity score $V_i = \max\{\hat{q}_{\alpha_{lo}}(X_i; \mathcal{Z}_{tr}) - T_i, T_i - \hat{q}_{\alpha_{hi}}(X_i; \mathcal{Z}_{tr})\}$
2: For each $i \in \mathcal{I}_{ca2}$ with $\Delta_i = 1$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{tr})$
3: Calculate the normalized weights $\hat{p}_i = \frac{W_i}{\sum_{i \in \mathcal{I}_{ca}} W_i + \hat{w}(X'; \mathcal{Z}_{tr})}$ and $\hat{p}_\infty = \frac{\hat{w}(X'; \mathcal{Z}_{tr})}{\sum_{i \in \mathcal{I}_{ca}} W_i + \hat{w}(X'; \mathcal{Z}_{tr})}$
4: Calculate η as the $(1-\alpha)$-th quantile of the distribution $\sum_{i \in \mathcal{I}_{ca}} \hat{p}_i \delta_{V_i} + \hat{p}_\infty \delta_\infty$
Output: $\hat{C}^2(X') = [\hat{q}_{\alpha_{lo}}(X'; \mathcal{Z}_{tr}, \mathcal{Z}_{ca1}) - \eta, \hat{q}_{\alpha_{hi}}(X'; \mathcal{Z}_{tr}, \mathcal{Z}_{ca1}) + \eta]$.

5. T-SCI: Improved Estimation

In the previous section, we propose WCCI, which has lower bound coverage guarantees. However, for better data efficiency, we want not only lower bound but also upper bounds. To reach the goal, we propose a two-stage algorithm T-SCI, which returns nearly perfect coverage.

The intuition is from Lemma 2, stating that the quantile conformal inference (QCI) returns a nearly perfect guarantee. Inspired by Lemma 2, we first use Algorithm 1 to return a temporal confidence band and then apply Quantile Conformal Inference to modify this band. We summarize the whole algorithm in Algorithm 2.

Remark 1. To apply T-SCI in practice, we split the dataset into $\mathcal{Z}_{tr}$, $\mathcal{Z}_{ca1}$ and $\mathcal{Z}_{ca2}$. We first apply Algorithm 1 with $\mathcal{Z}_{tr}$ and $\mathcal{Z}_{ca1}$ and return a confidence band. We then do calibration in Algorithm 2 with $\mathcal{Z}_{ca2}$.

Remark 2. In Algorithm 2, we modify the conformal score to be the absolute form again for clarity, since we have already derived an interval estimator of T in Algorithm 1.

Note that there are two conformal inference procedures in Algorithm 2. When the context is clear, the weight function $w(x)$ and the non-conformity score V_i refers to those in the second conformal inference. Before stating the theorem, we denote $H(X)$ to measure how well the quantile estimators $\hat{q}_{\alpha_{lo}}(X), \hat{q}_{\alpha_{hi}}(X)$ are.

$$H(X) = \max\{|\hat{q}_{\alpha_{lo}}(X) - q_{\alpha_{lo}}(X)|, |\hat{q}_{\alpha_{hi}}(X) - q_{\alpha_{hi}}(X)|\} \tag{7}$$

We next show in Theorem 5.1 that Algorithm 2 returns a guaranteed coverage either the weight or the temporal confidence band is estimated well.

Theorem 5.1 (Lower Bound) *Let $\hat{w}(x)$ be an estimate of the weight $w(x)$, $\hat{q}_{\alpha_{lo}}(x), \hat{q}_{\alpha_{hi}}(x)$ be the quantile estimator returned by WCCI, and $H(X)$ be defined as Equation 7. Assume that $\mathbb{E}[\hat{w}(X)|\mathcal{Z}_{tr}] = 1$ and $\mathbb{E}[w(X)] = 1$, where all the expectation operators $\mathbb{E}$ are taken over $X \sim \mathcal{P}_{X|\Delta=1}$. Denote $\hat{C}_n^2(x)$ as the output band of Algorithm 2 with n calibration samples, and denote X' as the testing point.*

From the weight perspective, under assumptions (A1):

A1. $\mathbb{E}_{X|\Delta=1}|\hat{w}(X) - w(X)| \le M_1$,
we have:

$$\lim_{n \to \infty} \mathbb{P}\left(T' \in \hat{C}_n^2(X')\right) \ge 1 - \alpha - \frac{1}{2}M_1.$$

From the quantile perspective, under assumptions (B1-B3):

B1. *$H(X) \le M_2$ a.s. w.r.t. X;*
B2. *There exists $\delta > 0$ such that $\mathbb{E}\hat{w}(X)^{1+\delta} < \infty$;*
B3. *There exists $\gamma, b_1, b_2 > 0$ such that $\mathbb{P}(T = t|X = x) \in [b_1, b_2]$ uniformly over all (x,t) with $t \in [q_{\alpha_{lo}}(x) - 2M_2 - 2\gamma, q_{\alpha_{lo}}(x) + 2M_2 + 2\gamma] \cup [q_{\alpha_{hi}}(x) - 2M_2 - 2\gamma, q_{\alpha_{hi}}(x) + 2M_2 + 2\gamma]$,*
we have:

$$\lim_{n \to \infty} \mathbb{P}\left(T' \in \hat{C}_n^2(X')\right) \ge 1 - \alpha - b_2(2M_2 + \gamma) - \frac{16M_2}{(M_2 + \gamma)^2 b_1}.$$

Theorem 5.1 demonstrates that when $M_1 \to 0$ *or* $M_2, \gamma \to 0$ as $|\mathcal{Z}_{tr}| \to 0$, the confidence band has guaranteed coverage with lower bound $1 - \alpha$. Compared to Theorem 4.1, Theorem 5.1 is doubly robust since the coverage is guaranteed when either (A1) or (B1-B3) holds. We next prove the upper bound in Theorem 5.2.

Theorem 5.2 (Upper Bound) *Let $\hat{w}(x)$ be an estimate of the weight $w(x)$, $\hat{q}_{\alpha_{lo}}(x), \hat{q}_{\alpha_{hi}}(x)$ be the quantile estimator returned by WCCI, and $H(X)$ be defined as Equation 7. Assume that $\mathbb{E}[\hat{w}(X)|\mathcal{Z}_{tr}] = 1$ and $\mathbb{E}[w(X)] = 1$, where all the expectation operators $\mathbb{E}$ are taken over $X \sim \mathcal{P}_{X|\Delta=1}$. Let $F_V \triangleq \sum_{i \in \mathcal{I}_{ca}} p_i \delta_{V_i} + p_\infty \delta_\infty$ be CDF, and assume V_i has no ties. Denote $\hat{C}_n^2(x)$ as the output band of Algorithm 2 with n calibration samples, and X' as the testing point. Under assumptions (C1-C4):*

C1. *$\forall \mathcal{S} \subset \mathcal{I}_{ca2}$, $|\sum_{i \in \mathcal{S}}(w(X_i) - \hat{w}(X_i))| \le M_1'$;*
C2. *$H(X) \le M_2'$ a.s. w.r.t. X;*
C3. *$F_V(t + L) - F_V(t) \ge KL$ for all t, L;*
C4. *there exists $b_1, b_2 > 0$ such that $\mathbb{P}(T = t|X =$*

[3]We use the two-sided band here for generality, and the one-sided band directly follows.

Table 1. Performance of Different Models on RRNLNPH.

Method	Total		Censored		Uncensored		Interval Length	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Cox Reg.	0.832	0.008	/	/	/	/	22.08	0.23
Random Survival Forest (Ishwaran et al., 2008)	0.948	0.006	/	/	/	/	16.39	0.22
Nnet-Survival (Gensheimer & Narasimhan, 2019)	0.834	0.007	0.560	0.016	0.982	0.005	20.11	0.37
MTLR (Yu et al., 2011)	0.830	0.008	0.554	0.017	0.980	0.003	19.85	0.34
CoxPH (Katzman et al., 2018)	0.829	0.008	0.554	0.016	0.978	0.004	19,65	0.21
CoxCC (Kvamme et al., 2019)	0.830	0.008	0.556	0.016	0.975	0.003	20.17	0.24
CoxPH+WCCI	0.912	0.03	0.854	0.047	0.949	0.028	21.59	0.90
CoxPH+T-SCI	0.974	0.009	0.947	0.018	0.994	0.005	29.85	0.68
CoxCC+WCCI	0.919	0.03	0.862	0.043	0.955	0.026	21.55	1.27
CoxCC+T-SCI	0.974	0.009	0.946	0.017	0.995	0.004	29.62	0.57
CoxPH+WCCI(unweigted)	0.907	0.020	0.830	0.049	0.949	0.006	22.06	1.27
CoxPH+T-SCI(unweighted)	0.950	0.018	0.875	0.049	0.990	0.012	27.72	3.13
CoxCC+WCCI(unweigted)	0.941	0.029	0.815	0.029	0.948	0.009	22.57	0.70
CoxCC+T-SCI(unweighted)	0.955	0.020	0.877	0.048	0.992	0.007	28.92	1.04
Kernel (Chen, 2020)	0.951	0.024	0.858	0.093	0.993	0.014	51.63	32.92

$x) \in [b_1, b_2]$ uniformly over all (x,t) with $t \in [q_{\alpha_{lo}}(x) - r, q_{\alpha_{lo}}(x) + r] \cup [q_{\alpha_{hi}}(x) - r, q_{\alpha_{hi}}(x) + r]$, where $r = 2M_2' + 2M_1'/K$ We have

$$\lim_{n\to\infty} \mathbb{P}\left(T' \in \hat{C}_n^2(X')\right) \le 1 - \alpha + b_2(2M_2' + \frac{M_1'}{K}).$$

Theorem 5.2 demonstrates that when the weight function $\hat{w}(x)$ and the quantile function $\hat{q}_{\alpha_{lo}}(x), \hat{q}_{\alpha_{hi}}(x)$ are estimated well (C1-C2), the returned coverage T-SCI has a lower bound guarantee. Combining Theorem 5.1 and Theorem 5.2 leads to a “nearly perfect” guaranteed coverage for T-SCI.

6. Experiments

This section aims at verifying some key arguments: (1) T-SCI returns valid coverage with small length interval; (2) Weight plays an essential role in the algorithm; (3) Censoring is more challenging than uncensoring settings. The results support these arguments both in synthetic data and real-world data, see Table 1 and 2.

6.1. Setup

Datasets and Environment. We conduct extensive experiments to test the efficiency of the algorithm. We use one synthetic dataset RRNLNPH (from Kvamme et al. (2019)) and two real-world datasets, METABRIC and SUPPORT (See Appendix B for more details). For each dataset, we test several baseline algorithms along with our proposed algorithms. In each run, 80% data are randomly sampled as the training data, and the two halves of the rest are randomly split as calibration data and test data. We run the experiments 100 times for each algorithm. Moreover, we collect the results under different αs.

Algorithms. We choose the linear Cox regression (labeled as Cox Reg.) as a baseline. Besides, we conduct CoxPH (Katzman et al., 2018) and CoxCC (Kvamme et al., 2019) (both belong to Cox-MLP) using $S(t)$ to return confidence band although they do not contain a theoretical guarantee. We also choose a kernel-based non-Cox method, (Chen, 2020) (labeled as Kernel). Besides, we conduct RSF (Ishwaran et al., 2008) (labeled as Random Survival Forest), Nnet-Survival (Gensheimer & Narasimhan, 2019) (labeled as Nnet-Survival), and MTLR (Yu et al., 2011) (labeled as MTLR) as benchmarks. We emphasize that these methods are different approaches since our method is Cox-based.

We integrate our proposed algorithms WCCI and T-SCI with CoxPH and CoxCC, labeled as CoxPH/CoxCC + WCCI/T-SCI. Besides, to ensure that the weights in WCCI and T-SCI are helpful, we test the *unweighted version* of WCCI and T-SCI, where we set all weights to 1 in WCCI and T-SCI. We summarize all the experimental results under significance level $\alpha = 95\%$ in Table 1. Ideally, a perfect method returns coverage slightly larger to 95% with small interval length while being balanced in censored and uncensored data.

Metrics. In the synthetic dataset, our core metric is *empirical coverage* (EC), defined as the fraction of testing points whose survival time falls in the predicted confidence band. Besides, we calculate the *average interval length* of the returned band. Given confidence level α, a perfect confidence band is expected to return empirical coverage larger than $1 - \alpha$ with a small interval length. In the real-world dataset, we use *surrogate empirical coverage* (SEC) instead of empirical coverage due to the lack of survival time. SEC

Table 2. Model Comparison on SUPPORT

Method	Total		Censored		Uncensored		Interval Length	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Cox Reg.	0.999	0.001	/	/	/	/	2026.68	1.03
Random Survival Forest (Ishwaran et al., 2008)	0.981	0.003	/	/	/	/	1890.11	10.75
Nnet-Survival (Gensheimer & Narasimhan, 2019)	0.995	0.002	0.988	0.005	0.998	0.001	1928.57	15.58
MTLR (Yu et al., 2011)	0.994	0.003	0.986	0.006	0.998	0.002	1921.65	37.69
CoxPH (Katzman et al., 2018)	0.981	0.002	0.992	0.004	0.975	0.002	2029.00	0.10
CoxCC (Katzman et al., 2018)	0.981	0.002	0.992	0.004	0.975	0.003	2553.32	9.43
CoxPH+WCCI	0.982	0.007	0.838	0.032	0.984	0.006	1940.85	16.41
CoxPH+T-SCI	0.993	0.004	0.923	0.026	0.993	0.003	2001.08	10.09
CoxCC+WCCI	0.980	0.008	0.829	0.047	0.983	0.008	1941.50	16.24
CoxCC+T-SCI	0.992	0.005	0.916	0.034	0.992	0.004	2001.69	10.67
CoxPH+WCCI(unweigted)	0.825	0.023	0.543	0.054	0.957	0.010	988.06	50.67
CoxPH+T-SCI(unweighted)	0.944	0.018	0.831	0.054	0.996	0.005	1712.76	83.19
CoxCC+WCCI(unweigted)	0.823	0.020	0.539	0.042	0.956	0.009	994.91	62.72
CoxCC+T-SCI(unweighted)	0.942	0.021	0.825	0.063	0.997	0.004	1705.41	75.66
Kernel (Chen, 2020)	0.988	0.020	0.936	0.178	0.988	0.038	2027.70	30.27

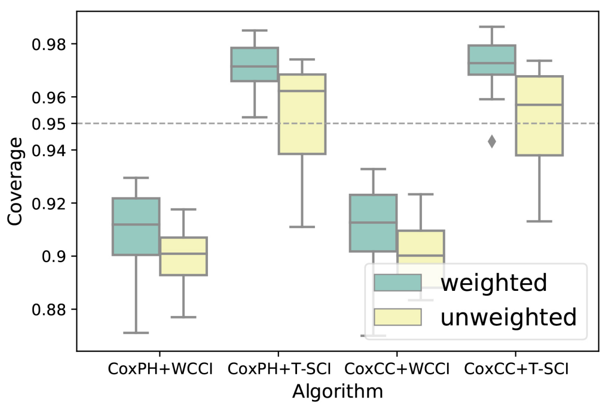


Figure 4. **Weight Rationality.** We compare the weighted version (green) with its corresponding unweighted version (yellow). The weighted versions contain less bias for WCCI (closer to the expected 95%) and less variance for T-SCI (shorter boxes).

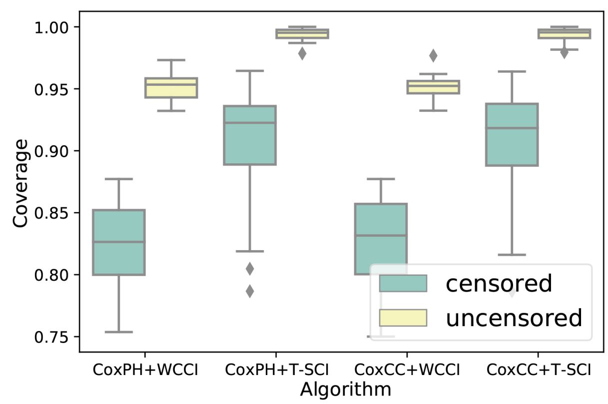


Figure 5. **Censoring comparison.** We compare model performances on censored (green) and uncensored (uncensored) data separately. All the algorithms show a larger coverage and less variance on censored data.

calculates the number when the censoring time is no larger than the band's upper bound for censored data, which is the upper bound of EC. Formally, let $\mathcal{I}_{test}$ be the training set index and $\hat{C}(X_i) = [T^l(X_i), T^u(X_i)]$ be the returned confidence band.. When the test set have the survival time T_i, namely, (X_i, T_i) in the test set, its empirical coverage (EC) is defined as:

$$EC \triangleq \frac{1}{|\mathcal{I}_{test}|} \sum_{i \in \mathcal{I}_{test}} \mathbb{I}(T_i \in \hat{C}(X_i)).$$

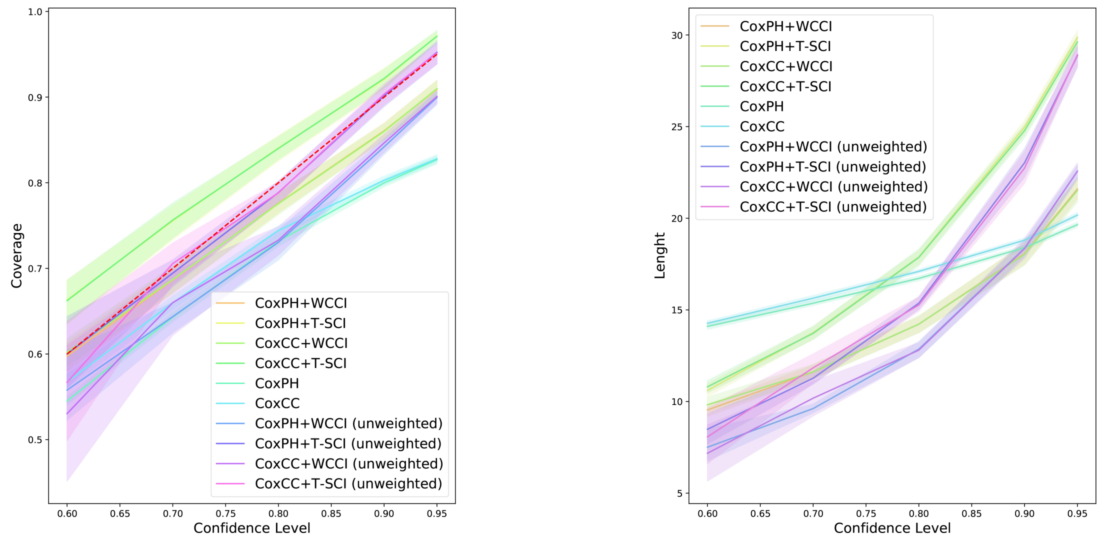
When the test set have only the observed time Y_i, namely, (X_i, Y_i, Δ_i) in the test set, its surrogate empirical coverage (SEC) is defined as:

$$SEC \triangleq \frac{1}{|\mathcal{I}_{test}|} \sum_{i \in \mathcal{I}_{test}} \mathbb{I}(Y_i \in \hat{C}(X_i), \Delta_i = 1) + \sum_{i \in \mathcal{I}_{test}} \mathbb{I}(Y_i \leq T^u(X_i), \Delta_i = 0)$$

[4]The survival time of censored data is accessible in the synthetic dataset.

6.2. Analysis

We summarize the experimental results of RRNLNPH in Table 1 and results of SUPPORT in Table 2, and we defer the results on METABRIC to Appendix B. Our analysis is mainly based on synthetic data RRNLNPH (the experiments on real-world datasets show similar trends).



(a) Coverage Comparison

(b) Interval Length Comparison

Figure 6. **Comparison under different confidence level ($1-\alpha$).** CoxCC+T-SCI (purple) and CoxPH+T-SCI (grey) returns guaranteed coverage under different confidence level without much increase of interval length (CoxCC+T-SCI and CoxPH+T-SCI overlap).

Coverage and Interval Length. Figure 6 shows the empirical coverage and the interval length of the confidence band returned by algorithms under different confidence levels. Ideally, the confidence band should have large coverage with a small interval length. Notice that WCCI and T-SCI based algorithms outperform their original versions, showing that the proposed algorithms work well. Furthermore, T-SCI is more conservative than WCCI, where T-SCI has larger coverage and interval length. We further remark that T-SCI returns guaranteed coverage (larger than $1-\alpha$) under different confidence levels.

The rationality of weight $w(x)$. We show the comparison between weighted and unweighted versions in Figure 4. For WCCI algorithms, the weighted version has coverage closer to $1-\alpha$. For T-SCI algorithms, the weighted version has lower variance. These results show that weighted versions outperform the unweighted versions, which validates the importance of weight $w(x)$.

The difficulty in censoring. Figure 5 shows the algorithms' performance on censored and uncensored data, respectively. All the algorithms perform larger coverage and less variance on the uncensored data, meaning that censored data is more challenging to deal with than uncensored data. Besides, we emphasize that although the unweighted versions may be closer to 95% in some cases, they lack theoretical guarantees and are imbalanced, meaning that it performs pretty differently on censored and uncensored data.

Analysis of Table 1. We show RRNLNPH results in Table 1, and results of SUPPORT and METABRIC perform similarly. Firstly, notice that WCCI does not reach the expected coverage mainly due to the inaccurate estimation on the weight, while T-SCI reaches it due to milder requirements in Theorem 5.1 (double robustness). Secondly, notice that Cox Reg., CoxPH, CoxCC (they all lack theoretical guarantee) all fail to return the proper coverage. Thirdly, unweighted versions all suffer from poor performances on censored data due to the lack of weight, showing that weight is vital in covariate shift. Finally, we emphasize that the Kernel method suffers from large interval length and large variance despite the moderate coverage. As a comparison, WCCI and T-SCI often perform more stably.

7. Conclusion

In this paper, we derive confidence band for Cox-based models. We first introduce WCCI by proposing a new nonconformity score. We then propose T-SCI, a two-stage conformal inference applying WCCI as input. Theoretical analysis shows that T-SCI returns nearly perfect coverage, meaning both lower and upper bound guarantee. We conduct extensive experiments on both synthetic data and real-world data to show the proposed algorithm's correctness.

Acknowledgements

This work has been partially supported by National Key R&D Program of China (2019AAA0105200), Zhongguancun Haihua Institute for Frontier Information Technology and the Institute Guo Qiang, Tsinghua University (2019GQG1002). We gratefully thank Lihua Lei, Jianhao Ma, Yuhao Wang and Dinghuai Zhang for their kind and constructive comments.

References

Akritas, M. G., Murphy, S. A., and Lavalley, M. P. The theil-sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association*, 90(429):170–177, 1995.

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019a.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*, 2019b.

Bellotti, T. and Crook, J. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(12):1699–1707, 2009.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 175–197, 2020.

Bostr, H., Asker, L., Gurung, R., Karlsson, I., Lindgren, T., Papapetrou, P., et al. Conformal prediction using random survival forests. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 812–817. IEEE, 2017.

Boström, H., Johansson, U., and Vesterberg, A. Predicting with confidence from survival data. In *Conformal and Probabilistic Prediction and Applications*, pp. 123–141, 2019.

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

Candès, E. J., Lei, L., and Ren, Z. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Chen, G. H. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, pp. 537–565. PMLR, 2020.

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

Feldmann, R. Leo-py: Estimating likelihoods for correlated, censored, and uncertain data with given marginal distributions. *Astronomy and Computing*, 29:100331, 2019.

Gensheimer, M. F. and Narasimhan, B. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

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

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

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18 (1):1–12, 2018.

Klein, J. P. and Moeschberger, M. L. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Kvamme, H., Borgan, Ø., and Scheel, I. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

Lee, C., Zame, W., Yoon, J., and van der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

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

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

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

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

Nakagawa, S. and Freckleton, R. P. Missing inaction: the dangers of ignoring missing data. *Trends in ecology & evolution*, 23(11):592–596, 2008.

Nouretdinov, I., Costafreda, S. G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., and Fu, C. H. Machine learning classification with confidence: application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression. *Neuroimage*, 56(2):809–813, 2011.

Robins, J. M. and Finkelstein, D. M. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

Romano, Y., Patterson, E., and Candès, E. J. Conformalized quantile regression. *arXiv preprint arXiv:1905.03222*, 2019.

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

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

Sargent, D. J. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91(S8):1636–1642, 2001.

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

Tibshirani, R. and Foygel, R. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 2019.

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

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

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

Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., and Azen, S. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24:1845–1853, 2011.

Zhu, R. and Kosorok, M. R. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.