CONFORMALIZED SURVIVAL ANALYSIS FOR GENERAL RIGHT-CENSORED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a framework to quantify predictive uncertainty in survival analysis, providing a reliable lower predictive bound (LPB) for the true, unknown patient survival time. Recently, conformal prediction has been used to construct such valid LPBs for *type-I right-censored data*, with the guarantee that the bound holds with high probability. Crucially, under the type-I setting, the censoring time is observed for all data points. As such, informative LPBs can be constructed by framing the calibration as an estimation task with covariate shift, relying on the conditionally independent censoring assumption. This paper expands the conformal toolbox for survival analysis, with the goal of handling the ubiquitous *general right-censored setting*, in which either the censoring or survival time is observed, but not both. The key challenge here is that the calibration cannot be directly formulated as a covariate shift problem anymore. Yet, we show how to construct LPBs with distribution-free finite-sample guarantees, under the same assumptions as conformal approaches for type-I censored data. Experiments demonstrate the informativeness and validity of our methods in simulated settings and showcase their practical utility on multi-modal breast cancer data.

1 INTRODUCTION

Survival analysis is essential in numerous fields, including medicine (Cole & Hudgens, 2010; Selvin, 2008), engineering (Ma & Krings, 2008), and social sciences (Cloyes et al., 2010). The primary objective is to predict the survival time T—the time-to-event such as death, failure, or relapse—based on covariates X that may include clinical markers, machine specifications, or sociological factors. The underlying challenge in survival analysis is that T is not observed for all subjects due to censoring. In particular, in this paper, we focus on *right-censored data*, the most common setting in survival analysis, where the survival event may not occur by the end of the follow-up period. This means that for some subjects X, the true survival time T is obscured by the censoring time C. As such, the observed censored survival time is given by $\tilde{T} = \min(T, C)$.

To formalize the above data generation process, we assume the triplets (X_i, T_i, C_i), $i = 1, \dots, N$, are sampled i.i.d. from $P_{X,T,C}$. The observed data, however, include censored observations, where we consider two possible forms of right-censored survival data.

Type-I right censoring: Here, we assume that censoring time C is observed for all subjects, which typically occurs in studies with a fixed duration, where all subjects are followed until the study ends. Under this setting, the observed dataset is of the form $\mathcal{D}_{\text{type-I}} = \{(X_i, \tilde{T}_i, C_i)\}_{i=1}^N$.

General right censoring: Unlike the type-I setting, in this more general case, we observe either C or T for each subject, but not both. Denote by $e_i = \mathbb{I}\{T_i < C_i\}$ a binary indicator variable that gets the value 1 if T_i is observed; that is, $\tilde{T} = T$ when $e = 1$ and $\tilde{T} = C$ otherwise. The observed dataset is then given by $\mathcal{D}_{\text{general}} = \{(X_i, \tilde{T}_i, e_i)\}_{i=1}^N$. Due to its greater generality, this setup is applicable for many survival analysis problems in biomedical research (Klein, 2003; Cole & Hudgens, 2010) and other fields. Thus, it is the focus of this work.

Given the observed censored data, the learning task is to estimate the distribution of $T \mid X$. However, since we do not fully observe the survival time T for all subjects, we need to impose further assumptions to make the learning task feasible. Indeed, a common assumption in survival analysis is that, conditional on the covariates X, the survival time T and the censoring time C are independent. Formally, this is expressed as follows.

Assumption 1.1 (Conditionally Independent Censoring (Kalbfleisch & Prentice, 2011))**.**

$$C \perp T \mid X.$$

Building on the above assumption and related ones, various machine learning approaches have been successful in estimating the distribution of $T|X$ (Cox, 1972; Wei, 1992; Lee et al., 2018; Nagpal et al., 2021b; Katzman et al., 2018). Despite their notable successes, these methods often lack reliability guarantees for the resulting predictions. This is attributed either to the opaque nature of modern deep learning algorithms or to the simplified modeling assumptions of more traditional statistical methods. Yet, such guarantees are of great importance given the harsh consequences of making erroneous predictions in high-stakes domains, such as healthcare (Navarro et al., 2021; Obermeyer et al., 2019).

1.1 The need for reliable lower predictive bounds for survival analysis

In response to this challenge, we aim to construct a lower predictive bound (LPB) on the survival time, supporting any given predictive model with distribution-free finite-sample guarantees (Vovk et al., 2005; Vovk, 2012; Tibshirani et al., 2019; Romano et al., 2019; Candès et al., 2023). Our goal is to construct an LPB function $\hat{L}(\cdot)$ such that, for a new test instance X_{test} with an unknown true survival time T_{test}, the event $\hat{L}(X_{\text{test}}) \leq T_{\text{test}}$ occurs with at least $1-\alpha$ probability. We refer to the probability of this event as the *coverage rate* of the LPB. For example, if the LPB satisfies the coverage property at level $(1-\alpha) = 0.9$, then $\hat{L}(X_{\text{test}})$ is guaranteed to be lower than the unknown T_{test} at least 90% of the time. To further enhance reliability, we seek to obtain a coverage guarantee that holds *conditional on the observed dataset*, in a manner similar to Gui et al. (2024); Park et al. (2019); Bates et al. (2021); Angelopoulos et al. (2021). To this end, let $\delta \in (0,1)$ be a tolerance level that accounts for the randomness in the realization of the dataset, and define the probably approximately correct (PAC)-type LPB as follows.

Definition 1.1. *Let (X_i, T_i, C_i), $i = 1, \dots, N$ be i.i.d. samples from $\mathbb{P}_{X,T,C}$, with $\hat{L}$ being a function of the observed dataset $\mathcal{D} = \mathcal{D}_{\text{general}} = \{(X_i, \tilde{T}_i, e_i)_{i=1}^N\}$, where $\tilde{T}_i = \min(T_i, C_i)$ and $e_i = \mathbb{I}[T_i < C_i]$. $\hat{L}$ is a PAC-type LPB at level $\alpha \in (0,1)$ with tolerance $\delta \in (0,1)$ if, with probability at least $1-\delta$ over the realization of $\mathcal{D}$,*

$$\mathbb{P}_{(X_{\text{test}}, T_{\text{test}}) \sim P}\big(T_{\text{test}} \geq \hat{L}(X_{\text{test}}) \mid \mathcal{D}\big) \geq 1-\alpha$$

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

Recently, conformal prediction methods have been applied to construct valid LPBs for *type-I right-censored data*. The key idea is to use holdout samples to calibrate the predictions of a survival analysis model, ensuring the desired coverage level is obtained for new test points. Under the assumption of conditionally independent censoring, Candès et al. (2023) introduced a conformal method that constructs valid and informative LPBs in finite samples. This approach was further refined by Gui et al. (2024), offering more informative LPBs with PAC-type guarantees. Both methods rely on the availability of the censoring time C for all subjects, which is used to discard early-censored samples. Intuitively, discarding such samples brings the observed time $\tilde{T}$ closer to the true event time T, yielding more accurate LPBs. Specifically, leveraging Assumption 1.1, Candès et al. (2023); Gui et al. (2024) reformulate the LPB construction as an estimation problem under covariate shift. Notably, the above calibration methods produce LPBs that are doubly robust, meaning they remain valid if either the predictive model or the estimation of the covariate shift likelihood ratio is accurate.

In the *general right-censored setting*, however, the literature is less developed and existing methods lack such strong validity guarantees. This is primarily because the censoring time C is not available for all subjects, making the calibration process more challenging.

1.2 Our Contributions

Building on the foundations of Gui et al. (2024), we propose a doubly robust calibration framework for constructing finite-sample valid LPBs for general right-censored survival analysis data.

To produce informative LPBs, we leverage the event indicator e and introduce a novel approach to selectively discard certain censored examples in a data-driven manner, bringing the censored event time $\tilde{T}$ closer to the true event time T. *Crucially, although the calibration cannot be directly formulated as a covariate shift problem, our approach forms valid LPBs under the same assumptions of conformal methods for the type-I setting.* We validate our theory and assess the applicability of the proposed methods using synthetic data, and further demonstrate their effectiveness on real breast cancer data. A Python implementation of our methods is provided in the supplementary material.

Concurrently with our work, Qin et al. (2024) and Meixide et al. (2024) developed distribution-free methods to create LPBs in this general right-censored setting. However, their bootstrap-based approaches are merely supported by asymptotic validity guarantees that hold under certain regularity conditions. This stands in contrast with the finite-sample guarantees provided by our methods. Furthermore, our novel approach of selectively discarding censored examples in a data-driven manner opens a new avenue for survival analysis under general right-censored data, diverging from the approaches by Qin et al. (2024) and Meixide et al. (2024).

2 Background and related work

Denote by $q_\tau(x)$ the true τ-th quantile of $T \mid X = x$, and define the oracle LPB for T_{test} as the α-th conditional quantile function, i.e., $L(X_{\text{test}}; \alpha) = q_\alpha(X_{\text{test}})$. While this oracle LPB is guaranteed to attain $1 - \alpha$ coverage by definition, in practice we do not have access to the true $q_\alpha(x)$, rendering this approach infeasible. As a way out, one could use survival analysis tools to estimate the conditional quantile function, which we denote by $\hat{q}_\tau(x)$, and form a heuristic LPB for T_{test} by setting $\hat{L}(X_{\text{test}}; \alpha) = \hat{q}_\alpha(X_{\text{test}})$. This heuristic approach, however, may not attain the desired coverage, unless the quantile estimates are accurate.

In this paper, we alleviate this issue by building on conformal prediction—a general framework to calibrate the estimated LPB. The appeal of conformal prediction is that it allows us to work with any predictive model and guarantee the desired $1-\alpha$ coverage at test time. The key idea is to modify the LPB function $\hat{L}(x; \tau)$ by rigorously tuning a hyperparameter $\tau \in \mathbb{R}$ using holdout calibration data $\mathcal{I}_{\text{cal}}$ to attain the desired coverage rate. We denote the tuned hyper-parameter as $\hat{\tau}$, and the resulting LPB as $\hat{L}(X; \hat{\tau})$.

2.1 First steps: naive conformalized survival analysis

Before presenting our method, we first outline a naive conformal prediction approach for tuning τ. In addition to using this approach as a baseline method in our experiments, it introduces the core principles of LPB calibration and motivates the need for more powerful solutions.

The key observation here is that $\tilde{T} \leq T$ by definition, and therefore a valid LPB for $\tilde{T}$ is also a valid (but conservative) LPB for T. As such, a naive approach to tune $\hat{\tau}$ is to discard the information provided by the event indicator e_i and construct a valid LPB for $\tilde{T}$ using the calibration points $\{(X_i, \tilde{T}_i)\}_{i \in \mathcal{I}_{\text{cal}}}$. In more detail, we define the LPB function as $\hat{L}(X; \tau) = \hat{q}_\tau(X)$, which is tightly connected to the methods by Chernozhukov et al. (2021); Gui et al. (2024).[1] Then, we formulate a (conservative) empirical estimator of the true miscoverage rate $\alpha(\tau) = \mathbb{P}(T < \hat{q}_\tau(X))$ for each choice of $\hat{q}_\tau(\cdot),\ \tau \in [0, 1]$, defined as follows:

$$\hat{\alpha}_{\text{naive}}(\tau) = \frac{1}{|\mathcal{I}_{\text{cal}}|} \sum_{i \in \mathcal{I}_{\text{cal}}} \mathbb{I}\left\{ \tilde{T}_i < \hat{q}_\tau(X_i) \right\}. \tag{1}$$

Above, we use the indicator function $\mathbb{I}(\cdot)$ to count the events where $\tilde{T}_i < \hat{q}_\tau(X_i)$, and so $\hat{\alpha}_{\text{naive}}(\tau)$ is an unbiased estimator of $\alpha_{\text{naive}}(\tau) = \mathbb{P}(\tilde{T} < \hat{q}_\tau(X))$. Since $\tilde{T} \leq T$ we get that $\alpha_{\text{naive}}(\tau) \geq \alpha(\tau)$, implying that the empirical quantity $\hat{\alpha}_{\text{naive}}(\tau)$ is anticipated to overestimate the miscoverage rate, on average. Finally, we define $\hat{\tau}_{\text{naive}}$ as the smallest value of τ for which $\hat{\alpha}_{\text{naive}}(\tau)$ meets the user-defined miscoverage level α, i.e.,

$$\hat{\tau}_{\text{naive}} = \sup\left\{ \tau \in [0, 1] : \sup_{\tau' \leq \tau} \hat{\alpha}_{\text{naive}}(\tau') \leq \alpha \right\}. \tag{2}$$

[1] There are other possible design choices for $\hat{L}(X; \tau)$, such as the conformalized quantile regression (CQR) approach (Romano et al., 2019) used by Candès et al. (2023).

In turn, the resulting calibrated LPB is given by $\hat{L}(X;\hat{\tau}_{\text{naive}}) = \hat{q}_{\hat{\tau}_{\text{naive}}}(X)$. Notably, this $\hat{L}(X;\hat{\tau}_{\text{naive}})$ is a PAC-type LPB on $\tilde{T}$, and thus also a PAC-type LPB on T. While this result is not novel, for completeness, we present Theorem A.1 in the Appendix, which proves its validity.

2.2 Utilizing type-I censored data for powerful calibration

Recall that $\hat{\alpha}_{\text{naive}}$ from Equation 1 is an overly conservative miscoverage estimator, particularly when early censoring occurs. This is because $\hat{\alpha}_{\text{naive}}$ counts the events in which either $C_i < \hat{q}_\tau(X_i)$ or $T_i < \hat{q}_\tau(X_i)$ hold. To form a less conservative estimator, Gui et al. (2024) suggest discarding the samples for which $C_i < \hat{q}_\tau(X_i)$, a strategy that is inline with Candès et al. (2023). Under Assumption 1.1, it can be shown that such a selection rule induces a covariate shift. In turn, Gui et al. (2024) introduced the following miscoverage estimator:

$$\hat{\alpha}_{\text{type-I}}(\tau) = \frac{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\tau^{\text{type-I}}(X_i)\cdot \mathbb{I}\{\hat{q}_\tau(X_i)\le C_i\}\cdot \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}}{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\tau^{\text{type-I}}(X_i)\cdot \mathbb{I}\{\hat{q}_\tau(X_i)\le C_i\}},$$

where the weights $\hat{w}_\tau^{\text{type-I}}(x)$, approximating $1/\mathbb{P}\left(\hat{q}_\tau(X)\le C \mid X = x\right)$, rigorously account for the covariate shift induced by the selection rule $\mathbb{I}\{\hat{q}_\tau(X_i) \le C_i\}$. Notably, $\hat{\alpha}_{\text{type-I}}(\tau)$ differs from $\hat{\alpha}_{\text{naive}}(\tau)$ from Equation 1 in that the latter uses all the samples for calibration, i.e., the selection rule indicator is always set to 1. Armed with $\hat{\alpha}_{\text{type-I}}(\tau)$, the hyper-parameter $\hat{\tau}_{\text{type-I}}$ can be tuned analogously to Equation 2. Finally, the LPB of a new test point is given by $\hat{L}(X;\hat{\tau}_{\text{type-I}}) = \hat{q}_{\hat{\tau}_{\text{type-I}}}(X)$.

The advantage of the estimator $\hat{\alpha}_{\text{type-I}}(\tau)$ is that it is mostly less conservative compared to $\hat{\alpha}_{\text{naive}}(\tau)$, resulting in more informative LPBs. However, for *general right-censored data*, using $\hat{\alpha}_{\text{type-I}}(\tau)$ is not feasible as the censorship time C_i is not available for some subjects. Our work addresses this knowledge gap, showing how to rigorously utilize the event indicator $e_i = \mathbb{I}\{T_i < C_i\}$ to construct not only valid but also useful LPBs, as detailed in the next section.

3 Proposed methods

An intuitive adaptation of the miscoverage estimator proposed by Gui et al. (2024) to *general right-censored data* would be to examine only calibration points for which $e_i = 1$. The challenge with this selection rule is that the distribution shift it induces cannot be formulated as a covariate shift. This is because, even under Assumption 1.1, we expect statistical dependence between T and e conditional on X. Furthermore, without careful adjustments, estimating the miscoverage using only uncensored subjects jeopardizes the LPB's validity as the survival times of uncensored subjects are generally lower than the ones of the general population.

Nevertheless, in this section, we show that with a proper weighting scheme, the above set of ideas can be formalized into a valid calibration method. Concretely, we propose two calibration techniques. Our first method, called `focused` calibration, only uses uncensored subjects for LPB construction. We prove the validity of this method and show that, while it can lead to more informative LPBs than the `naive` approach, it remains conservative to some extent. This observation drives the proposal of our second approach, which builds upon and further improves the first. Our key idea is to reduce conservativeness by including certain censored examples—in addition to the uncensored subjects—when calibrating the LPB. We term this method `fused` calibration, as it utilizes both censored and uncensored samples while maintaining the theoretical guarantees.

3.1 Focused calibration

In what follows, we formally introduce our `focused` calibration method, then provide an intuitive explanation for its correctness, and finally characterize the condition under which it leads to LPBs that are more informative than those produced by the `naive` method. Following Gui et al. (2024), we use the LPB function $\hat{L}(X;\tau) = \hat{q}_\tau(X)$, but we tune the hyper-parameter τ by utilizing the special structure of the general right-censored setting.

In more detail, we calibrate the LPB using samples with $e_i = 1$, i.e., those for which $\tilde{T}_i = T_i$. With this selection rule in place, we define the resulting miscoverage estimator as

$$\hat{\alpha}_{\text{focus}}(\tau) = \frac{\sum_{i \in \mathcal{I}_{\text{cal}}} \hat{w}(X_i) \cdot \mathbb{I}\{e_i = 1\} \cdot \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}}{\sum_{i \in \mathcal{I}_{\text{cal}}} \hat{w}(X_i) \cdot \mathbb{I}\{e_i = 1\}}, \tag{3}$$

where the weights $\hat{w}(x)$ approximate $1/\mathbb{P}\,(e = 1 \mid X = x)$. Later, we explain the rationale behind the use of $\hat{w}(x)$ in detail. For now, we remark that: (i) the estimation of $\mathbb{P}(e = 1 \mid X = x)$ can be done by fitting any binary classifier to the training pairs (X, e); and (ii) in Section 3.3, we further characterize the effect of the estimation error on the resulting coverage.

Next, the tuned $\hat{\tau}_{\text{focus}}$ is defined as the smallest τ such that $\hat{\alpha}_{\text{focus}}(\tau)$ passes the user-specified miscoverage level α, i.e.

$$\hat{\tau}_{\text{focus}} = \sup\left\{\tau \in [0, 1] : \sup_{\tau' \leq \tau} \hat{\alpha}_{\text{focus}}(\tau') \leq \alpha\right\}.$$

Finally, the LPB for a new test point is given by $L(X_{\text{test}}; \hat{\tau}_{\text{focus}}) = \hat{q}_{\hat{\tau}_{\text{focus}}}(X_{\text{test}})$. For ease of reference, the algorithm summarizing our `focused` calibration procedure is in Appendix B.3.

Building upon the foundations of Gui et al. (2024), we establish a double robustness result for the validity of our method. This result implies that the LPBs are approximately valid if either (i) the weights $\hat{w}(x)$ are well approximated; or (ii) the estimated quantiles $\hat{q}_\tau(x)$, $\tau \in (0, 1)$ of the conditional distribution $T \mid X$ are accurate. Additionally, if (ii) is satisfied, the LPBs are approximately valid conditional on X_{test}. The theorems formalizing the validity of the `focused` method is postponed to Section 3.3, as this procedure is actually a specific instance of the `fused` calibration method, which is presented in the next section.

Having presented the algorithm, we turn to discuss why the use of the oracle weight $w(x) = 1/\mathbb{P}(e = 1 \mid X = x)$, estimated by $\hat{w}(x)$, conservatively accounts for the distribution shift induced by the selection rule from Equation 3. Our analysis follows Gui et al. (2024), providing an upper bound for the true miscoverage rate $\alpha(\tau)$ obtained by the quantile estimator $\hat{q}_\tau$ for some $\tau \in (0, 1)$:

$$\begin{aligned}
\alpha(\tau) = \mathbb{P}\big(T < \hat{q}_\tau(X)\big) &= \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X) \mid X)\right] && (4)\\
&= \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X) \mid X) \cdot \mathbb{P}(e = 1 \mid X) \cdot w(X)\right] \\
&\leq \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X), e = 1 \mid X) \cdot w(X)\right] && (5)\\
&= \mathbb{E}\left[\mathbb{I}\left\{T < \hat{q}_\tau(X), e = 1\right\} \cdot w(X)\right] && (6)\\
&= \mathbb{E}\left[\mathbb{I}\{\tilde{T} < \hat{q}_\tau(X)\} \cdot \mathbb{I}\left\{e = 1\right\} \cdot w(X)\right] = \alpha_{\text{focus}}(\tau). && (7)
\end{aligned}$$

Above, steps 4 and 6 hold by the tower property, and the inequality in step 5 is due to Lemma B.1, provided in the Appendix. Importantly, the above derivation shows that, while we do not have direct access to the true $\alpha(\tau)$, we can upper bound it using a weighted average of observed quantities, as revealed by step 7. In turn, the miscoverage upper bound $\alpha_{\text{focus}}(\tau)$ provides the basis for constructing its empirical estimator $\hat{\alpha}_{\text{focus}}(\tau)$ from Equation 3, with the denominator serving to normalize the weighted average.

While being conservative, our experiments show that the LPBs derived from the `focused` method tend to be more informative than those produced by the `naive` approach from Section 2.1. To better understand when this is the case, we now present the condition under which $\alpha_{\text{focus}}(\tau)$ is less conservative than $\alpha_{\text{naive}}(\tau)$.

Proposition 3.1. *Under Assumption 1.1, the relation $\alpha_{\text{focus}}(\tau) < \alpha_{\text{naive}}(\tau)$ holds when*

$$\mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 1) < \mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 0) \quad \forall x \in \mathcal{X}. \tag{8}$$

The proof is given in Appendix B.2. To highlight the practical implications of the above proposition, we recall the definition of $\tilde{T} = \min(T, C)$ and e, and re-write Condition 8 as follows:

$$\mathbb{P}(T < \hat{q}_\tau(X) \mid X = x, e = 1) < \mathbb{P}(C < \hat{q}_\tau(X) \mid X = x, e = 0) \quad \forall x \in \mathcal{X}.$$

As such, our `focused` method is anticipated to outperform the `naive` approach when the probability of early survival time for subjects with $e = 1$ is smaller than the probability of early censorship time for subjects with $e = 0$. Such a phenomenon can occur in medical trials, where early censoring is prevalent due to factors such as non-compliance (Zhou et al., 2020), withdrawal of consent (Wilson et al., 2021), and loss to follow-up (Fontana et al., 2018; Gill et al., 2018; Monfared et al., 2021).

With that said, Proposition 3.1 also reveals that the `naive` approach from Section 2.1 may produce tighter LPBs than the `focused` method. Specifically, this occurs for samples that do not satisfy the inequality in Equation 8. This insight naturally brings the idea of detecting such samples and leveraging them to mitigate the conservativeness of $\hat{\alpha}_{\text{focus}}(\tau)$. Indeed, this is the key principle behind our `fused` calibration method, described hereafter.

3.2 Fused calibration

Following the above discussion, we show how to fuse the `naive` calibration approach with the `focused` calibration method to enhance statistical efficiency, making better use of available data. In addition to using all the uncensored points for calibration, we aim to selectively include in the calibration set the censored subjects that violate the inequality in Equation 8. The oracle selection criterion is thus formally stated as

$$s_\tau(x) = \begin{cases} 1, & \text{if } \mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 0) < \mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 1), \\ 0, & \text{otherwise.} \end{cases}$$

Observe that we cannot compute $s_\tau(x)$ in practice as we do not have access to the conditional distribution $\tilde{T} \mid X, e$. However, this indicator $s_\tau(x)$ can be estimated using a single classifier, as follows. First, train a binary classifier with (X_i, e_i) as input, predicting the label $\mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}$. Then, use this classifier to compare the estimated probabilities of the label given $(X = x, e = 0)$ and $(X = x, e = 1)$. We refer to the fitted estimator of $s_\tau(x)$ as $\hat{s}_\tau(x)$. Crucially, inaccurate estimation of $\hat{s}_\tau$ does not impact the validity of the calibration procedure, however it can affect the gain in statistical efficiency.

With $\hat{s}_\tau(x)$ in place, we formulate the fused selection rule,

$$\zeta_\tau(X_i, e_i) = \begin{cases} 1, & \text{if } \hat{s}_\tau(X_i) = 1, \\ e_i, & \text{otherwise,} \end{cases} \tag{9}$$

which uses the calibration point $(X_i, \tilde{T}_i)$ for miscoverage estimation if either $e_i = 1$ or $\hat{s}_\tau(X_i) = 1$. Next, the fused miscoverage estimator is formulated as follows:

$$\hat{\alpha}_{\text{fused}}(\tau) = \frac{\sum_{i \in \mathcal{I}_{\text{cal}}} \hat{w}_\tau(X_i) \cdot \mathbb{I}\{\zeta_\tau(X_i, e_i) = 1\} \cdot \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}}{\sum_{i \in \mathcal{I}_{\text{cal}}} \hat{w}_\tau(X_i) \cdot \mathbb{I}\{\zeta_\tau(X_i, e_i) = 1\}},$$

where the weights $\hat{w}_\tau(x)$ approximate $w_\tau(x) = 1/\mathbb{P}(\zeta_\tau(x, e) = 1 \mid X = x)$. We note that since $\hat{s}_\tau(X_i)$ is deterministic given X_i, we have that

$$w_\tau(X_i) = \begin{cases} 1, & \text{if } \hat{s}_\tau(X_i) = 1, \\ 1/\mathbb{P}(e_i = 1 \mid X_i), & \text{otherwise.} \end{cases}$$

Continuing with the same rationale as in `focused` calibration, we define $\hat{\tau}_{\text{fused}}$ as

$$\hat{\tau}_{\text{fused}} = \sup\left\{\tau \in [0,1] : \sup_{\tau' \le \tau} \hat{\alpha}_{\text{fused}}(\tau') \le \alpha\right\},$$

and set the LPB for a new test point as $\hat{L}(X_{\text{test}}; \hat{\tau}_{\text{fused}}) = \hat{q}_{\hat{\tau}_{\text{fused}}}(X_{\text{test}})$.

For ease of reference, the `fused` calibration method is presented in Algorithm 1. Notably, we use training data (independent from the calibration points) to fit the estimated quantile functions $\hat{q}_\tau(x)$, the classifiers $\hat{s}_\tau(x)$, and the weights $\hat{w}_\tau(x)$; these models serve as inputs to the `fused` calibration algorithm. We also remark that Algorithm 1 reduces to the `focused` approach with the choice of $\hat{s}_\tau(x) := 0$ for all x, and to the `naive` approach when $\hat{s}_\tau(x) := 1$ for all x. For this reason, we argue that the `fused` method can be viewed as an interpolation between two valid calibration strategies, with the optimal interpolation hyperparameter being estimated by $\hat{s}_\tau(x)$. This perspective, as well as the derivations in Appendix C.1 support the intuition for the method's validity, formally proven next.

Algorithm 1 Conformalized survival analysis for general right-censored data: `fused` calibration

Input: desired miscoverage level α; calibration data $\mathcal{D}_{\text{cal}} = \{(X_i, \tilde{T}_i, e_i)\}_{i\in\mathcal{I}_{\text{cal}}}$; estimated quantiles of $T \mid X$, $\{\hat{q}_\tau(\cdot)\}_{\tau\in[0,1]}$; weights $\hat{w}_\tau(x)$ approximating $1/\mathbb{P}(\zeta_\tau(X,e) = 1 \mid X = x)$, with the corresponding selection rule $\zeta_\tau : \mathcal{X} \times \{0,1\} \to \{0,1\}$ defined in Equation 9.

Procedure:

1: **for** τ in a grid over $[0,1]$ **do**[2]

2: Compute the fused miscoverage estimator:

$$\hat{\alpha}_{\text{fused}}(\tau) = \frac{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\tau(X_i)\cdot \mathbb{I}\{\zeta_\tau(X_i,e_i)=1\}\cdot \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}}{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\tau(X_i)\cdot \mathbb{I}\{\zeta_\tau(X_i,e_i)=1\}}$$

3: **end for**

4: Calibrate the hyperparameter τ:

$$\hat{\tau}_{\text{fused}} = \sup\left\{\tau\in[0,1] : \sup_{\tau'\le\tau} \hat{\alpha}_{\text{fused}}(\tau') \le \alpha\right\}$$

5: **Return:** The calibrated LPB: $\hat{L}(\cdot) := \hat{L}(\cdot;\hat{\tau}_{\text{fused}}) = \hat{q}_{\hat{\tau}_{\text{fused}}}(\cdot)$

3.3 Theoretical guarantees

Our theoretical framework builds on the work by Gui et al. (2024) and adapts their double robustness result to our proposed methods for general right-censored data. Concretely, our first validity result states that if the weights are well-estimated, i.e., $\hat{w}_\tau(x) \approx w_\tau(x)$, then the resulting coverage is approximately higher than $1-\alpha$ conditional on the training data $\mathcal{D}_{\text{tr}} = \{(X_i,\tilde{T}_i,e_i)\}_{i\in\mathcal{I}_{\text{tr}}}$ and the holdout calibration data $\mathcal{D}_{\text{cal}} = \{(X_i,\tilde{T}_i,e_i)\}_{i\in\mathcal{I}_{\text{cal}}}$, where $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{cal}}$. In particular, this guarantee holds even when the estimated conditional quantiles $\hat{q}_\tau$ are inaccurate. Further, following Theorem 3.1 below, we can see that the guaranteed coverage rate gets closer to the desired level as (i) $\hat{w}_\tau$ gets closer to w_τ; and (ii) the size of the calibration set $|\mathcal{I}_{\text{cal}}|$ increases.

Theorem 3.1 (Approximate calibration with accurate weights)**.** *Fix any $\alpha,\delta\in(0,1)$. Given that $\hat{q}_\tau(x)$ is non-decreasing and continuous in τ, and that there exists some constant $\hat{\gamma}_\tau > 0$ such that $\hat{w}_\tau(x) \le \hat{\gamma}_\tau$ for P_X-almost all x. Under Assumption 1.1, with probability at least $1-\delta$ over the draw of $\mathcal{D}$, the LPB $\hat{L}$ produced by either Algorithm 3 or by Algorithm 1 satisfies*

$$\mathbb{P}\left[T \ge \hat{L}(X)|\mathcal{D}\right] \ge 1-\alpha-\Delta_w,$$

where

$$\Delta_w := \sup_{\tau\in[0,1]}\left(\mathbb{E}\left[\left|\frac{\hat{w}_\tau(X_i)}{w_\tau(X_i)\,\pi_\tau} - 1\right|\middle|\mathcal{D}_{\text{tr}}\right] + \sqrt{\frac{1+\frac{\hat{\gamma}_\tau^2}{\pi_\tau^2} + \max\left(1, \frac{\hat{\gamma}_\tau^2}{\pi_\tau^2}-1\right)^2}{0.4\,|\mathcal{I}_{\text{cal}}|}\ln\frac{1}{\delta}}\right),$$

and $\pi_\tau := \mathbb{E}_{X\sim P_X}\left[\frac{\hat{w}_\tau(X)}{w_\tau(X)}|\mathcal{D}_{\text{tr}}\right]$.

The proof is given in Appendix C.2.

The next result states a stronger, approximated conditional coverage guarantee that holds even when the weights are inaccurate. In essence, this result holds under a stronger assumption that the conditional quantiles are well-estimated, i.e., $\hat{q}_\tau(x) \approx q_\tau(x)$. This is stated formally in assumption (b). Before formalizing this intuition, we define the smallest estimated quantile level to bound T with probability at least β, as

$$\tau(\beta) = \sup\left\{\tau'\in[0,1] : \mathbb{P}\left(T < \hat{q}_{\tau'}(X) \mid \mathcal{D}_{\text{tr}}\right) \le \beta\right\}.$$

Theorem 3.2 (Approximate calibration with accurate conditional quantiles)**.** *Assume the same conditions as Theorem 3.1 and assume further that the distribution of $T \mid X$ is continuous and its density is upper bounded by a constant $B > 0$, and that there exists a constant $r > 0$ such that*

[2] Observe that the values of τ that lead to a change in $\hat{\alpha}(\tau)$ correspond to those for which $\tilde{T}_i = \hat{q}_\tau(X_i)$, for $i\in\mathcal{I}_{\text{cal}}$. Consequently, the grid is defined over the τ values that satisfy this equality, with $\tau = 0$ included.

(a) *For P_X-almost all x:* $\sup_{\xi \in [\tau(\alpha), \tau(\alpha+r)+\psi]} w_\xi(x) \leq \gamma$ *and* $\sup_{\xi \in [\tau(\alpha), \tau(\alpha+r)+\psi]} \hat{w}_\xi(x) \leq \hat{\gamma}$ *for some constants* $\gamma, \hat{\gamma}, \psi \geq 0$

(b) $\sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ \max(B,1) |\hat{q}_{\tau(\beta)}(x) - q_\beta(x)| + \hat{\gamma}\gamma \sqrt{\frac{\log(1/\delta)}{0.4|\mathcal{I}_{\text{cal}}|}} \right\} \leq r$, *where* $q_\beta(x)$ *is the* β*-quantile of the distribution of* $T \mid X = x$.

Then, under Assumption 1.1, with probability at least $1-\delta$ *over the draw of* $\mathcal{D}$*, the LPB* $\hat{L}$ *produced by either Algorithm 3 or by Algorithm 1 satisfies that for* P_X*-almost all* x*,*

$$\mathbb{P}_{(X,T)\sim P}\left[T \geq \hat{L}(X) | \mathcal{D}, X = x\right] \geq 1 - \alpha - \Delta_q,$$

where $\Delta_q = \sup_{\beta \in [\alpha, \alpha+r]} \sup_{x \in \mathcal{X}} \left\{ 2B \cdot \left|\hat{q}_{\tau(\beta)}(x) - q_\beta(x)\right| \right\} + \hat{\gamma}\gamma \sqrt{\frac{1}{0.4|\mathcal{I}_{\text{cal}}|} \log \frac{1}{\delta}}$.

The proof is deferred to Appendix C.3.

4 Experiments

4.1 Simulation studies

We follow the experimental protocol outlined by Candès et al. (2023) and Gui et al. (2024) to evaluate the performance of our methods across various simulated scenarios. Notably, since real-world data lack the true survival time T for all test points, synthetic data experiments are important to verify the correctness of our theory.

Base predictive models In all experiments, we utilize the *DeepSurv* method (Katzman et al., 2018), as the base predictor to estimate the conditional quantiles $\hat{q}_\tau$. We also fit Random Forest classifiers to estimate the weights $\hat{w}_\tau$ and indicator $\hat{s}_\tau$ for our calibration methods. For more details, see Appendix D.

Data We test our calibration methods on six simulated settings, where the covariates X are drawn from a uniform distribution, $X \sim U[0,4]^p$. The conditional survival time $T \mid X$ follows a lognormal distribution, as in Candès et al. (2023); Gui et al. (2024):

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

where $\mathcal{N}(\mu(X), \sigma^2(X))$ is the normal distribution with mean $\mu(X)$ and standard deviation $\sigma(X)$ defined in Table 1 for each setting, along with the dimension p of X. The distribution of $C \mid X$ is composed of (i) a base censorship distribution $C_{\text{base}}(X)$, defined in Table 1 for each setting; and (ii) day-one censorship, simulating a realistic scenario where a small proportion of subjects experience immediate censorship. These individuals are selected randomly from the top 20% of the population based on their first covariate $X(0)$, capturing the heterogeneous nature of early censorship (Zhou et al., 2020; Wilson et al., 2021). Formally, we define

$$C \mid X = x \sim \begin{cases} C_{\text{base}}(x), & \text{with probability } 1 - 0.2 \cdot \mathbb{I}\{x(0) > 3.2\}, \\ 0, & \text{with probability } 0.2 \cdot \mathbb{I}\{x(0) > 3.2\}. \end{cases}$$

Methods and performance metrics Since we are the first to design calibration methods that achieve finite-sample PAC-style LPBs for general right-censored data, we compare our `focused` and `fused` algorithms only to the `naive` method from Section 2, which is supported by similar guarantees. In all simulation studies, we set the desired coverage level $1-\alpha$ to 90% and evaluate both the coverage rate and the average LPB over the test set.

Results The performance metrics for the synthetic experiments are presented in Figure 1. Following the top panel in that figure, we can see that all calibration methods achieve a valid coverage rate across all settings, as expected. Notably, our `fused` method tends to be less conservative than both the `focused` and `naive` approaches. This empirical evidence supports the theoretical motivation behind the design of the `fused` method. To further stress the advantage of the latter, we highlight

Table 1: Parameters specifying the distribution of survival time $T|X$ and censoring time $C|X$ in the six simulated data sets.

Setting	p	$\mu(x)$	$\sigma(x)$	$C_{\text{base}}(x)$
1	1	$0.632x$	2	$\text{Exp}(0.1)$
2	1	$3 \cdot 1\{x > 2\} + x \cdot 1\{x \leq 2\}$	2	$\text{Exp}(0.1)$
3	1	$2 \cdot 1\{x > 2\} + x \cdot 1\{x \leq 2\}$	2	$\text{Exp}(0.02 + \frac{x}{25})$
4	1	$3 \cdot 1\{x > 2\} + 1.5x \cdot 1\{x \leq 2\}$	2	lognormal $\left(2 + (\frac{2-X}{50}), 0.5\right)$
5	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	2	$\text{Exp}\left(\frac{x_{10}}{40} + \frac{1}{20}\right)$
6	10	$0.126(x_1 + \sqrt{x_3 x_5}) + 1$	$\frac{x_2+2}{2}$	$\text{Exp}\left(\frac{x_{10}}{40} + \frac{1}{20}\right)$

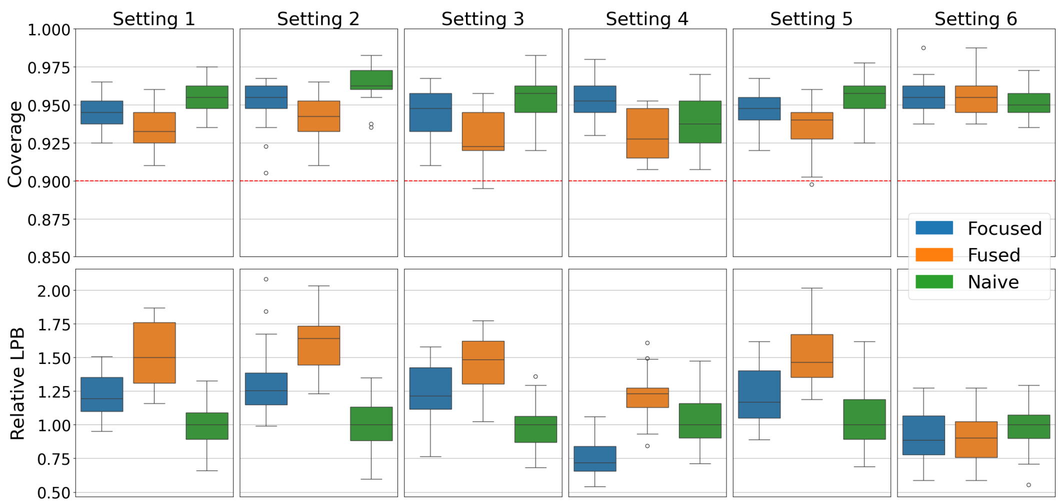

Figure 1: Performance of the different calibration methods for each of the six simulated settings from Table 1. **Top**: empirical coverage rate, with a red dashed line indicating the nominal 90% level. **Bottom**: relative LPB, defined as the LPB obtained by each method divided by the `naive` method's median LPB. A higher relative LPB is better. The reported performance metrics are evaluated on 25 independent trials, each consisting of newly sampled train, validation, calibration, and test sets of sizes 2400, 400, 800, and 400, respectively.

Setting 4, where one can see that (i) the `focused` method is more conservative than the `naive` approach, suggesting the extreme violation of Equation 8; and (ii) the `fused` method is less conservative than the other two calibration methods. Turning to LPB comparisons, the bottom panel of Figure 1 illustrates that the `fused` method generates the most informative lower bounds (higher is better). This statistical gain is in line with the tighter coverage rate of the `fused` method.

4.2 Application to Real Data

We demonstrate the practical utility of our methods by applying them to *The Cancer Genome Atlas Breast Invasive Carcinoma* (TCGA-BRCA) multi-modal dataset collection (Tomczak et al., 2015). We choose to work with this dataset for several key reasons. First, it holds significant scientific value for machine-aided prognostics by integrating diverse genomic, transcriptomic, and clinical data. Second, its multi-modal nature underscores the ability of our methods to process diverse, high-dimensional data. Third, this real-world dataset presents realistic, challenging survival data that includes subjects with early censorship—a common occurrence in medical studies. The challenge of constructing informative LPBs in the presence of early censorship is central to our work as well as previous contributions, such as those by Candès et al. (2023); Gui et al. (2024).

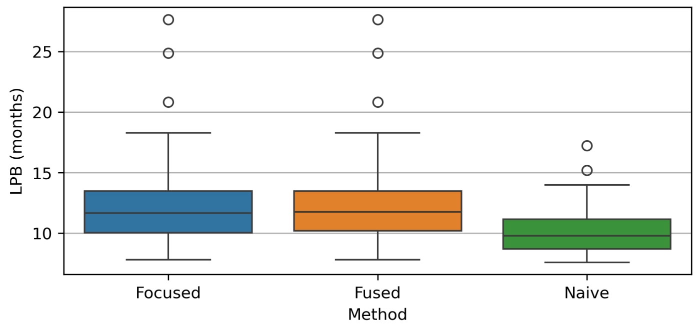

Figure 2: Average LPB obtained by the different calibration methods for the TCGA-BRCA dataset. Results are evaluated on 50 random splits of the data, each consisting of distinct train, validation, calibration, and test sets of sizes $626, 71, 276$, and 71, respectively.

Concretely, the TCGA-BRCA dataset includes various clinical features such as age, cancer grade, and tumor size, along with pathological biomarkers, genetic data, and visual biopsy samples. To effectively estimate the conditional survival times from such diverse features, we designed dedicated preprocessing and training procedures for this data. These include the use of the GigaPath foundation model (Xu et al., 2024) to extract features from the visual biopsy samples, extraction of cancer grade from patients' medical reports, and more. A detailed description of the entire processing pipeline can be found in Appendix D.

Figure 2 presents the average LPB values obtained by the `naive`, `focused`, and `fused` calibration methods. As shown, both the `focused` and `fused` methods yield higher LPBs compared to the `naive` approach, which aligns with the general trend observed in our synthetic experiments. To better understand the differences in performance, recall that the `focused` method is expected to outperform the `naive` approach when Condition 8 holds, which motivated the formulation of $\hat{s}_{\mathcal{T}}$. By measuring the mean value of $\hat{s}_{\hat{\tau}_{\text{fused}}}$ we can gain insight into the mechanism of the `fused` method as well as the degree to which Condition 8 holds. Specifically, we found the mean value of $\hat{s}_{\hat{\tau}_{\text{fused}}}$ to be very low for this data, being equal to 0.01. This suggests that Condition 8 approximately holds, explaining the superior performance of the `focused` method compared to the `naive` approach. Moreover, Equation 9 indicates that for such a small mean value of $\hat{s}_{\hat{\tau}_{\text{fused}}}$, the `focused` and `fused` methods should behave almost the same, as indeed happened in practice.

In contrast with the synthetic experiments, here it is not possible to rigorously verify that the desired coverage level is achieved in practice—we do not have access to ground truth survival times. Yet, our theory indicates that the `naive` method should attain a conservative coverage level. Further, under well-approximated estimation of the weights or quantiles, the proposed `focused` and `fused` methods should produce less conservative and approximately valid LPBs.

5 Discussion

We introduced a flexible uncertainty quantification framework with finite-sample LPB coverage guarantees for general right-censored survival analysis data. Numerical experiments confirmed the validity of our methods and showed that the `fused` method tends to outperform both the `focused` and `naive` approaches.

The validity of the proposed calibration methods relies on the conditionally independent censoring and i.i.d. assumptions, which may not hold in all real-world cases. It is therefore of great importance to study the effect of violations of these assumptions both to understand the practical limitations of our methods and to further enhance their robustness. In particular, the work by Oliveira et al. (2024) may serve as a valuable starting point to move beyond the i.i.d. assumption; and the work by Feldman & Romano (2024) can be further utilized to build robustness to imperfect training data.

References

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118.

Kristin G Cloyes, Bob Wong, Seth Latimer, and Jose Abarca. Time to prison return for offenders with serious mental illness released from prison: A survival analysis. *Criminal Justice and Behavior*, 37(2):175–187, 2010.

Stephen R Cole and Michael G Hudgens. Survival analysis in infectious disease research: describing events in time. *Aids*, 24(16):2423–2431, 2010.

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

Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clinical Cancer Research*, 14(16):5158–5165, 08 2008.

Shai Feldman and Yaniv Romano. Robust conformal prediction using privileged information, 2024.

Matteo Fontana, Luca Boeri, Andrea Gallioli, Elisa De Lorenzis, Franco Palmisano, Stefano Paolo Zanetti, Gianluca Sampogna, Giancarlo Albo, Fabrizio Longo, Franco Gadda, et al. Prevalence and predictors of being lost to follow-up after transurethral resection of the prostate. *Scientific reports*, 8(1):6406, 2018.

Amitoj Gill, Rohit Gosain, Shruti Bhandari, Rahul Gosain, Gurkirat Gill, Joseph Abraham, and Kenneth Miller. "lost to follow-up" among adult cancer survivors. *American Journal of Clinical Oncology*, 41(10):1024–1027, 2018.

Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111(2):459–477, 2024.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1.

Mert Ketenci, Shreyas Bhave, Noemie Elhadad, and Adler J. Perotte. Maximum likelihood estimation of flexible survival densities with importance sampling. In *MLHC*, pp. 360–380, 2023.

John P Klein. *Survival analysis: techniques for censored and truncated data*, pp. 70. Springer Science Business Media, Inc, second edition, 2003.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Zhanshan Ma and Axel W Krings. Survival analysis approach to reliability, survivability and prognostics and health management (phm). In *2008 IEEE Aerospace Conference*, pp. 1–20. IEEE, 2008.

Carlos García Meixide, Michael R. Kosorok, and Marcos Matabuena. Uncertainty quantification for intervals, 2024.

Sara Monfared, Anna Martin, Don Selzer, and Annabelle Butler. Travel distance reduces follow-up compliance but has no effect on long-term weight loss success in bariatric patients. *Surgical endoscopy*, 35:1579–1583, 2021.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021a.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep cox mixtures for survival regression. In *Machine Learning for Healthcare Conference*, pp. 674–708. PMLR, 2021b.

Constanza L. Andaur Navarro, Johanna Aag Damen, Toshihiko Takada, Steven W. J. Nijman, Paula Dhiman, Jie Ma, Gary Stephen Collins, Ram Chandra Bajpai, Richard D. Riley, Karel G. M. Moons, and Lotty Hooft. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *The BMJ*, 375, 2021.

Ziad Obermeyer, Brian W. Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447 – 453, 2019.

Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jing Qin, Jin Piao, Jing Ning, and Yu Shen. Conformal predictive intervals in survival analysis: a re-sampling approach, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Steve Selvin. *Survival analysis for epidemiologic and medical research*. Cambridge University Press, 2008.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology (Poznan, Poland)*, 19 (1A):A68–77, 2015. ISSN 1428-2526. doi: 10.5114/wo.2014.47136.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Lee-Jen Wei. 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.

Brooke Wilson, Michelle Nadler, Alexandra Desnoyers, and Eitan Amir. Quantifying withdrawal of consent, loss to follow-up, early drug discontinuation, and censoring in oncology trials. *Journal of the National Comprehensive Cancer Network : JNCCN*, 19:1–8, 09 2021. doi: 10.6004/jnccn.2021.7015.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

Brenda Zhou, Travis C Mitchell, Alexander M Rusakevich, David M Brown, and Charles C Wykoff. Noncompliance in prospective retina clinical trials: analysis of factors predicting loss to follow-up. *American journal of ophthalmology*, 210:86–96, 2020.

A NAIVE METHOD'S ALGORITHM AND PROOF OF VALIDITY

A formal description of the algorithm for the `naive` calibration method is given by Algorithm 2.

Algorithm 2 Conformalized survival analysis for general right-censored data: `naive` calibration

Input: desired miscoverage level α; calibration data $\mathcal{D}_{\text{cal}}$; estimated quantiles of $T \mid X$, $\{\hat{q}_\tau(\cdot)\}_{\tau\in[0,1]}$.

Procedure:

1: **for** τ in a grid over $[0,1]$ **do**

2: Compute the fused miscoverage estimator:

$$\hat{\alpha}_{\text{naive}}(\tau) = \frac{1}{|\mathcal{I}_{\text{cal}}|}\sum_{i\in\mathcal{I}_{\text{cal}}} \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}$$

3: **end for**

4: Calibrate the hyperparameter τ:

$$\hat{\tau}_{\text{naive}} = \sup\left\{\tau \in [0,1] : \sup_{\tau'\leq\tau} \hat{\alpha}_{\text{naive}}(\tau') \leq \alpha\right\}$$

5: **Return:** The calibrated LPB: $\hat{L}(\cdot) := \hat{L}(\cdot;\hat{\tau}_{\text{naive}}) = \hat{q}_{\hat{\tau}_{\text{naive}}}(\cdot)$

Theorem A.1 (PAC-type validity of the naive LPB)**.** *Fix any $\alpha, \delta \in (0,1)$. Assume that $\hat{q}_\tau(x)$ is non-decreasing in τ. With probability at least $1-\delta$ over the draw of $\mathcal{D}$, the LPB produced by the* `naive` *calibration method from satisfies*

$$\mathbb{P}\left[T \geq \hat{L}(X;\hat{\tau}_{\text{naive}})|\mathcal{D}\right] \geq 1-\alpha-\Delta_n,$$

where $\Delta_n := \sqrt{\frac{1}{|\mathcal{I}_{\text{cal}}|}\ln\frac{1}{\delta}}$.

Proof. For the proof we denote the smallest estimated quantile to bound T with probability at least β, conditional on the training data, as

$$\tau(\beta) = \sup\left\{\tau' \in [0,1] : \mathbb{P}\left(T < \hat{q}_{\tau'}(X) \mid \mathcal{D}_{\text{tr}}\right) \leq \beta\right\}.$$

As in the proof for theorem 3 in the work by Gui et al. (2024), proving that $1-\delta \leq \mathbb{P}\left[\hat{\tau}_{\text{naive}} \leq \tau\left(\alpha+\Delta_n\right) \mid \mathcal{D}_{\text{tr}}\right]$ is sufficient for proving the theorem. Analyzing the `naive` miscoverage estimator, we get that

$$\begin{aligned}
&\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)+\epsilon\right) \leq \alpha \mid \mathcal{D}_{\text{tr}}\right] \\
&= \mathbb{P}\left[\sum_{i\in\mathcal{I}_{\text{cal}}} \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\} \leq \left|\mathcal{I}_{\text{cal}}\right| \alpha \middle| \mathcal{D}_{\text{tr}}\right] \\
&= \mathbb{P}\left[\sum_{i\in\mathcal{I}_{\text{cal}}} \left(\mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\} - \alpha\right) \leq 0 \middle| \mathcal{D}_{\text{tr}}\right] \\
&\leq \mathbb{E}\left[t \exp\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \left(\alpha - \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\}\right)\right) \middle| \mathcal{D}_{\text{tr}}\right] \qquad (10)
\end{aligned}$$

where the first transition is by the definition of $\hat{\alpha}_{\text{naive}}$, and the last by the Markov inequality. Further conditioning on $\left(X_i\right)_{i\in\mathcal{I}_{\text{cal}}}$, since $p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\}$ is $\frac{1}{4}$-sub-gaussian, we can use Hoeffding's lemma we get that

$$\begin{aligned}
&\mathbb{E}\left[t \exp\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \left(p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\}\right)\right) \middle| \left(X_i\right)_{i\in\mathcal{I}_{\text{cal}}}, \mathcal{D}_{\text{tr}}\right] \\
&\leq \exp\left(t \sum_{i\in\mathcal{I}_{\text{cal}}} \mathbb{E}\left[p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\} \middle| \left(X_i\right)_{i\in\mathcal{I}_{\text{cal}}}, \mathcal{D}_{\text{tr}}\right] + \frac{t^2}{8} \sum_{i\in\mathcal{I}_{\text{cal}}} 1^2\right) \\
&= \exp\left(t \sum_{i\in\mathcal{I}_{\text{cal}}} \left(p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{P}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\} \middle| \left(X_i\right)_{i\in\mathcal{I}_{\text{cal}}}, \mathcal{D}_{\text{tr}}\right] + \frac{t^2\left|\mathcal{I}_{\text{cal}}\right|}{8}\right).
\end{aligned}$$

Since $\tilde{T}_i \leq T_i$ by definition, we have that

$$p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{P}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\} \leq 0,$$

and so by the law of total expectation, we get that

$$\begin{aligned}
&\mathbb{E}\left[t \exp\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \left(p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right) - \mathbb{I}\left\{\tilde{T}_i < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right\}\right)\right) \middle| \mathcal{D}_{\text{tr}}\right] \\
&\leq \exp\left(\frac{\left|\mathcal{I}_{\text{cal}}\right| t^2}{8}\right).
\end{aligned}$$

Combining with 10, we get that

$$\begin{aligned}
&\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)+\epsilon\right) \leq \alpha \mid \mathcal{D}_{\text{tr}}\right] \\
&\leq \exp\left(\frac{\left|\mathcal{I}_{\text{cal}}\right| t^2}{8}\right) \mathbb{E}\left[t \exp\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \left(\alpha - p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right)\right) \middle| \mathcal{D}_{\text{tr}}\right].
\end{aligned}$$

By the definition of $\tau\left(\alpha+\Delta_n\right)$, we have that

$$\mathbb{P}\left[T < \hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X\right) \mid \mathcal{D}_{\text{tr}}\right] \leq \alpha + \Delta_n,$$

and so,

$$
\begin{aligned}
&\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)+\epsilon\right)\leq\alpha\mid\mathcal{D}_{\text{tr}}\right] \\
&\leq \exp\left(\frac{\left|\mathcal{I}_{\text{cal}}\right|t^2}{8}\right)\mathbb{E}\left[t\exp\left(\sum_{i\in\mathcal{I}_{\text{cal}}}\left(\mathbb{P}\left[T<\hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\mid\mathcal{D}_{\text{tr}}\right]\right.\right.\right. \\
&\qquad\qquad\qquad\qquad \left.\left.\left.-\Delta_n-p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\right)\right)\middle|\mathcal{D}_{\text{tr}}\right] \\
&\leq \exp\left(\frac{\left|\mathcal{I}_{\text{cal}}\right|t^2}{8}-t\Delta_n\left|\mathcal{I}_{\text{cal}}\right|\right),
\end{aligned}
$$

where the last inequality follows from the $\frac{1}{4}$-sub-gaussianity of

$$p_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)-\mathbb{P}\left[T<\hat{q}_{\tau(\alpha+\Delta_n)+\epsilon}\left(X_i\right)\mid\mathcal{D}_{\text{tr}}\right].$$

Taking

$$t=4\Delta_n,$$

we get that

$$\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)+\epsilon\right)\leq\alpha\mid\mathcal{D}_{\text{tr}}\right]\leq\exp\left(-2\left|\mathcal{I}_{\text{cal}}\right|\left(\Delta_n\right)^2\right).$$

Since $\left(-2\left|\mathcal{I}_{\text{cal}}\right|\left(\Delta_n\right)^2\right)\leq-\frac{\left|\mathcal{I}_{\text{cal}}\right|\cdot\left(\mathbb{E}\left[\left|\frac{\hat{w}(X)}{w(X)}-\pi\right|\middle|\mathcal{D}_{\text{tr}}\right]-\pi\Delta_n\right)^2}{\pi^2+\hat{\gamma}^2+\tilde{\gamma}^2}$, we get that

$$\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)+\epsilon\right)\leq\alpha\mid\mathcal{D}_{\text{tr}}\right]\leq\delta,$$

where the last inequality follows from the definition of Δ_n. Taking the limit as $\epsilon\to 0$, we get that

$$1-\delta\leq\mathbb{P}\left[\hat{\alpha}_{\text{naive}}\left(\tau\left(\alpha+\Delta_n\right)\right)>\alpha\mid\mathcal{D}_{\text{tr}}\right]\leq\mathbb{P}\left[\hat{\tau}_{\text{naive}}<\tau\left(\alpha+\Delta_n\right)\mid\mathcal{D}_{\text{tr}}\right],$$

as needed. □

B FURTHER DETAILS REGARDING THE FOCUSED CALIBRATION METHOD

B.1 FOCUSED CALIBRATION LEMMA

Lemma B.1. *Suppose that $x\in\mathcal{X},q:\mathcal{X}\to\mathbb{R}^+$. If $\mathbb{P}(T<q(X)\mid X=x)>0$ and under Assumtion 1.1,*

$$\mathbb{P}\left(e=1,T<q(X)\,|X=x\right)\geq\mathbb{P}\left(e=1\,|X=x\right)\mathbb{P}\left(T<q(X)\,|X=x\right),$$

where the probability is taken over $P_{(X,C,T)}$.

Proof. Suppose $x\in\mathcal{X}$. All derivations in this proof are conditional on $X=x$, and so we denote $\mathbb{P}_x(\cdot)=\mathbb{P}(\cdot\mid X=x)$ and $q=q(X)$.

We begin by developing $\mathbb{P}_x(T<C\mid T<q)$:

$$
\begin{aligned}
\mathbb{P}_x(T<C\mid T<q)&=\mathbb{P}_x(T<C\mid T<q,C\geq q)\mathbb{P}_x(C\geq q)+\mathbb{P}_x(T<C\mid T<q,C<q)\mathbb{P}_x(C<q)\\
&=1\cdot\mathbb{P}_x(C\geq q)+\frac{\mathbb{P}_x(T<C,T<q,C<q)}{\mathbb{P}_x(T<q,C<q)}\mathbb{P}_x(C<q)\\
&=\mathbb{P}_x(C\geq q)+\frac{\mathbb{P}_x(T<C,T<q,C<q)}{\mathbb{P}_x(T<q)\cdot\mathbb{P}_x(C<q)}\mathbb{P}_x(C<q)\\
&=\mathbb{P}_x(C\geq q)+\frac{\mathbb{P}_x(T<C,T<q,C<q)}{\mathbb{P}_x(T<q)}\\
&=\mathbb{P}_x(C\geq q)+\frac{\mathbb{P}_x(T<C,C<q)}{\mathbb{P}_x(T<q)},
\end{aligned}
$$

where the first transition is by the law of total probability and the third is by Assumption 1.1.

We now develop $\mathbb{P}_x(T < C)$:

$$\mathbb{P}_x(T < C) = \mathbb{P}_x(T < C, C < q) + \mathbb{P}_x(T < C, C \geq q)$$

By subtraction the two quantities, we get

$$\begin{aligned}
&\mathbb{P}_x(T < C \mid T < q) - \mathbb{P}_x(T < C) \\
&\qquad = \mathbb{P}_x(C \geq q) + \frac{\mathbb{P}_x(T < C, C < q)}{\mathbb{P}_x(T < q)} - (\mathbb{P}_x(T < C, C < q) + \mathbb{P}_x(T < C, C \geq q)) \\
&\qquad = (\mathbb{P}_x(C \geq q) - \mathbb{P}_x(T < C, C \geq q)) + (\frac{\mathbb{P}_x(T < C, C < q)}{\mathbb{P}_x(T < q)} - \mathbb{P}_x(T < C, C < q)) \\
&\qquad = \mathbb{P}_x(T \geq C, C \geq q) + \left(\frac{1}{\mathbb{P}_x(T < q)} - 1\right)\mathbb{P}_x(T < C, C < q) \geq 0,
\end{aligned}$$

and so, multiplying by $\mathbb{P}_x(T < q)$ we get that

$$\mathbb{P}_x(T < C, T < q) - \mathbb{P}_x(T < C)\mathbb{P}_x(T < q) \geq 0.$$

That is,

$$\mathbb{P}_x(T < C, T < q) \geq \mathbb{P}_x(T < C)\mathbb{P}_x(T < q).$$

□

B.2 FOCUSED CALIBRATION INFORMATIVENESS CONDITION

In this section, we prove Proposition 3.1.

Proof. By looking at 5, we have that

$$\alpha_{\text{focus}}(\tau) = \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X), e = 1 \mid X) \cdot w(X)\right] = \mathbb{E}\left[\mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 1)\right].$$

Analogously, we can re-write $\alpha_{\text{naive}}(\tau)$ as

$$\alpha_{\text{naive}}(\tau) = \mathbb{P}(\tilde{T} < \hat{q}_\tau(X)) = \mathbb{E}\left[\mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x)\right].$$

Therefore, $\alpha_{\text{focus}}(\tau) < \alpha_{\text{naive}}(\tau)$ when

$$\mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x, e = 1) < \mathbb{P}(\tilde{T} < \hat{q}_\tau(X) \mid X = x) \quad \forall x \in \mathcal{X}.$$

□

B.3 FORMAL DESCRIPTION OF THE FOCUSED CALIBRATION ALGORITHM

A formal description of the `focused` calibration algorithm is given in Algorithm 3.

Note that the theorems for the validity of this algorithm in Section 3.3 are valid under the definition $\forall \tau \in [0, 1], w_\tau := w, \hat{w}_\tau := \hat{w}$.

C PROOFS FOR FUSED CALIBRATION DOUBLE ROBUSTNESS

C.1 FUSED CALIBRATION THEORETICAL DERIVATION AND LEMMA

The derivation for the `fused` calibration method is very similar to that of `focused` calibration, with $\zeta_\tau(X, e)$ replacing e:

Algorithm 3 Conformalized survival analysis for general right-censored data: `focused` calibration

Input: desired miscoverage level α; calibration data $\mathcal{D}_{\text{cal}}$; estimated quantiles of $T \mid X$, $\{\hat{q}_\tau(\cdot)\}_{\tau\in[0,1]}$; weights $\hat{w}(x)$ approximating $1/\mathbb{P}(e=1 \mid X=x)$.

Procedure:

1: **for** τ in a grid over $[0,1]$ **do**

2: Compute the fused miscoverage estimator:

$$\hat{\alpha}_{\text{focus}}(\tau) = \frac{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}(X_i)\cdot \mathbb{I}\{e_i=1\}\cdot \mathbb{I}\{\tilde{T}_i < \hat{q}_\tau(X_i)\}}{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}(X_i)\cdot \mathbb{I}\{e_i=1\}}$$

3: **end for**

4: Calibrate the hyperparameter τ:

$$\hat{\tau}_{\text{focus}} = \sup\left\{\tau\in[0,1] : \sup_{\tau'\leq\tau} \hat{\alpha}_{\text{focus}}(\tau') \leq \alpha\right\}$$

5: **Return:** The calibrated LPB: $\hat{L}(\cdot) := \hat{L}(\cdot;\hat{\tau}_{\text{focus}}) = \hat{q}_{\hat{\tau}_{\text{focus}}}(\cdot)$

$$\begin{aligned}
\alpha(\tau) = \mathbb{P}\big(T < \hat{q}_\tau(X)\big) &= \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X) \mid X)\right] && (11)\\
&= \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X) \mid X)\cdot \mathbb{P}(\zeta_\tau(X,e)=1 \mid X)\cdot w(X)\right]\\
&\leq \mathbb{E}\left[\mathbb{P}(T < \hat{q}_\tau(X), \zeta_\tau(X,e)=1 \mid X)\cdot w(X)\right] && (12)\\
&= \mathbb{E}\left[\mathbb{I}\left\{T < \hat{q}_\tau(X), \zeta_\tau(X,e)=1\right\}\cdot w(X)\right] && (13)\\
&= \mathbb{E}\left[\mathbb{I}\{\tilde{T} < \hat{q}_\tau(X)\}\cdot \mathbb{I}\left\{\zeta_\tau(X,e)=1\right\}\cdot w(X)\right] = \alpha_{\text{focus}}(\tau).
\end{aligned}$$

Where this time $w(x) = 1/\mathbb{P}(\zeta_\tau(X,e)=1 \mid X=x) = \min(1/\mathbb{P}(e=1 \mid X=x), s_\tau(X))$, steps Equation 11 and Equation 13 hold by the tower property, and step Equation 12 holds by the following lemma -

Lemma C.1. *Let $s:\mathcal{X}\to\{0,1\}$, and $\zeta_\tau:\mathcal{X}\times\{0,1\}\to\{0,1\}$ as defined in Equation 9. Then Under Assumtion 1.1, $\forall x\in\mathcal{X}, q:\mathcal{X}\to\mathbb{R}^+$,*

$$\mathbb{P}[\zeta_\tau(X,e)=1, \tilde{T}<q(X)\,|X=x] - \mathbb{P}[\zeta_\tau(X,e)=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x] \geq 0.$$

Proof. We'll prove this lemma by separating the cases into the two possible values of $\hat{s}_\tau(X)$, which is constant given $X=x$. Suppose that $x\in\mathcal{X}$ satisfies $\hat{s}_\tau(x)=0$. We have that

$$\begin{aligned}
&\mathbb{P}[\zeta_\tau(X,e)=1, \tilde{T}<q(X)\,|X=x]\\
&= \mathbb{P}[e=1, \tilde{T}<q(X)\,|X=x]\\
&= \mathbb{P}[e=1, T<q(X)\,|X=x],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}[\zeta_\tau(X,e)=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x]\\
&= \mathbb{P}[e=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x].
\end{aligned}$$

And so, by Lemma B.1, we have that

$$\begin{aligned}
&\mathbb{P}[\zeta_\tau(X,e)=1, \tilde{T}<q(X)\,|X=x] - \mathbb{P}[\zeta_\tau(X,e)=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x]\\
&= \mathbb{P}[e=1, T<q(X)\,|X=x] - \mathbb{P}[e=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x] \geq 0.
\end{aligned}$$

For $x\in\mathcal{X}$ with $\hat{s}_\tau(X)=1$, we have that

$$\begin{aligned}
&\mathbb{P}[\zeta_\tau(X,e)=1, \tilde{T}<q(X)\,|X=x]\\
&= \mathbb{P}[\tilde{T}<q(X)\,|X=x],
\end{aligned}$$

And

$$\begin{aligned}
&\mathbb{P}[\zeta_\tau(X,e)=1\,|X=x]\mathbb{P}[T<q(X)\,|X=x]\\
&= \mathbb{P}[T<q(X)\,|X=x].
\end{aligned}$$

And so, since by definition $\tilde{T} \leq T$, we have that $\forall x \in \mathcal{X}$:

$$\mathbb{P}[\zeta_\tau(X,e)=1, \tilde{T} < q(X) \,|\, X=x] - \mathbb{P}[\zeta_\tau(X,e)=1 \,|\, X=x]\mathbb{P}[T < q(X) \,|\, X=x]$$
$$= \mathbb{P}[\tilde{T} < q(X) \,|\, X=x] - \mathbb{P}[T < q(X) \,|\, X=x] \geq 0.$$

□

C.2 FUSED CALIBRATION WITH APPROXIMATELY ACCURATE WEIGHTS

In this section we prove Theorem 3.1, relying on the proof of (Gui et al., 2024, Theorem 3).

Proof. Here, we provide a single proof for the validity of both the `focused` and `fused`, using the appropriate notations for each method. For the `focused` method, we consider $\hat{\alpha}(\cdot) := \hat{\alpha}_{\text{focused}}(\cdot)$, $\hat{\tau} := \hat{\tau}_{\text{focused}}$, and $\zeta_\lambda(X_i, e_i) = e_i$. For the `fused` method, we denote $\hat{\alpha}(\cdot) := \hat{\alpha}_{\text{fused}}(\cdot)$, $\hat{\tau} := \hat{\tau}_{\text{fused}}$, and $\zeta_\lambda(X_i, e_i) = \mathbb{I}\{s_\lambda(X_i) = 1 \text{ or } e_i = 1\}$. Importantly, in both cases, the weights are defined as $w_\tau(x) := 1/\mathbb{P}(\zeta_\lambda(X,e) = 1 \mid X = x)$. We remark that all our claims hold for each of the two choices of these terms.

We define the error term by:

$$\Delta := \sup_{\lambda \in [0,1]} \left\{ \mathbb{E}\left[\left| \frac{\hat{w}_\lambda(X)}{w_\lambda(X)\pi_\lambda} - 1 \right| \middle| \mathcal{D}_{\text{tr}} \right] + \sqrt{\frac{1 + \frac{\hat{\gamma}_\lambda^2}{\pi_\lambda^2} + \max\left(1, \frac{\hat{\gamma}_\lambda}{\pi_\lambda} - 1\right)^2}{0.4|\mathcal{I}_{\text{cal}}|} \cdot \log\left(\frac{1}{\delta}\right)} \right\}$$

Recall that the oracle quantity is formulated as:

$$\tau(\alpha + \Delta) = \sup\left\{\lambda \in [0,1] : \mathbb{P}\left(T < \hat{q}_\lambda(X) \mid \mathcal{D}_{\text{tr}}\right) \leq \alpha + \Delta\right\}.$$

The outline of this proof builds on the proof of (Gui et al., 2024, Theorem 3). First, if $1-\delta \leq \mathbb{P}(\hat{\tau} \leq \tau(\alpha+\Delta) \mid \mathcal{D}_{\text{tr}})$, then the event $\{\hat{\tau} \leq \tau(\alpha+\Delta)\}$ holds with probability at least $1-\delta$. Therefore:

$$\begin{aligned} &\mathbb{P}(T \geq \hat{q}_{\hat{\tau}} \mid \mathcal{D}) \\ \geq &\mathbb{P}(T \geq \hat{q}_{\tau(\alpha+\Delta)}(X) \mid \mathcal{D}) \\ \geq &1 - \alpha - \Delta. \end{aligned}$$

Above, the first inequality is due to the monotonicity of $\hat{q}_\tau$ and the second one follows from the left-continuity of $\mathbb{P}(T \geq \hat{q}_\tau(X) \mid \mathcal{D})$ in τ. In what follows, we focus on showing $1-\delta \leq \mathbb{P}(\hat{\tau} \leq \tau(\alpha+\Delta) \mid \mathcal{D}_{\text{tr}})$. Suppose that $\varepsilon > 0$. For simplicity, we denote $\lambda := \tau(\alpha+\Delta)+\varepsilon$. Following the definition of $\hat{\alpha}(\tau)$, we get:

$$\begin{aligned} &\mathbb{P}(\hat{\alpha}(\tau(\alpha+\Delta)+\varepsilon) \leq \alpha \mid \mathcal{D}_{\text{tr}}) \\ =&\mathbb{P}(\hat{\alpha}(\lambda) \leq \alpha \mid \mathcal{D}_{\text{tr}}) \\ =&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\}}{\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)} \leq \alpha \middle| \mathcal{D}_{\text{tr}} \right) \\ =&\mathbb{P}\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\} \leq \alpha \sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i) \middle| \mathcal{D}_{\text{tr}} \right) \quad (14) \\ =&\mathbb{P}\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i) \left[\mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\} - \alpha\right] \leq 0 \middle| \mathcal{D}_{\text{tr}} \right) \\ =&\mathbb{P}\left(\sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i) \left[\alpha - \mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\}\right] \geq 0 \middle| \mathcal{D}_{\text{tr}} \right) \end{aligned}$$

Next, we apply Markov's inequality for any $t > 0$, and have:

$$\begin{aligned} 14 &\leq \mathbb{E}\left(\exp\left(t \sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i) \left[\alpha - \mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\}\right] \right) \middle| \mathcal{D}_{\text{tr}} \right) \\ &= \mathbb{E}\left(\exp\left(t \sum_{i\in\mathcal{I}_{\text{cal}}} \hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i) \left[\alpha + p_\lambda(X_i) - p_\lambda(X_i) - \mathbb{I}\{\tilde{T}_i < \hat{q}_\lambda(X_i)\}\right] \right) \middle| \mathcal{D}_{\text{tr}} \right). \end{aligned} \quad (15)$$

Above $p_\lambda(x) := \mathbb{P}(T < \hat{q}_\lambda(X) \mid X = x, \mathcal{D}_{\text{tr}})$. We now conditioning on $(X_i, e_i)_{i\in\mathcal{I}_{\text{cal}}}$

$$\begin{aligned}
&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
&\overset{(a)}{\leq}\exp\left(t\mathbb{E}\left(\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)+\frac{t^2\sum_{i\in\mathcal{I}_{\text{cal}}}(2\hat{w}_\lambda(X_i))^2}{8}\right)\\
&\overset{(b)}{\leq}\exp\left(t\mathbb{E}\left(\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)+\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right)\\
&\leq\exp\left(t\mathbb{E}\left(\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\left[\zeta_\lambda(X_i,e_i)p_\lambda(X_i)-\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)+\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right)\\
&\overset{(c)}{\leq}\exp\left(t\left(\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\mathbb{E}\left[\zeta_\lambda(X_i,e_i)p_\lambda(X_i)-\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\middle|\mathcal{D}_{\text{tr}},X_i,e_i\right]\right)+\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right),
\end{aligned}\tag{16}$$

where step (a) uses Hoeffding's inequality; step (b) follows from the boundness of $\hat{w}$; step (c) is due to the independence assumption between the samples. We now turn to develop the expectation inside the sum.

$$\begin{aligned}
&\mathbb{E}\left[\zeta_\lambda(X_i,e_i)p_\lambda(X_i)-\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\\
=&\mathbb{E}\left[\zeta_\lambda(X_i,e_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\mathbb{E}\left[p_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]-\mathbb{E}\left[\zeta_\lambda(X_i,e_i)\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\\
=&\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\mathbb{E}\left[p_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]-\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1,\tilde{T}_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\\
=&\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\mathbb{E}\left[p_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]-\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1,\tilde{T}_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\\
=&\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\mathbb{P}(T_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i)\\
&-\mathbb{P}\left(\zeta_\lambda(X_i,e_i)=1\mid\mathcal{D}_{\text{tr}},X_i,e_i\right)\mathbb{P}\left(\tilde{T}_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right)\\
=&\mathbb{P}\left[\zeta_\lambda(X_i,e_i)=1\mid\mathcal{D}_{\text{tr}},X_i,e_i\right]\left[\mathbb{P}(T_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i)-\mathbb{P}\left(\tilde{T}_i<\hat{q}_\lambda(X_i)\mid\mathcal{D}_{\text{tr}},X_i,e_i\right)\right]\\
&\leq 0
\end{aligned}$$

The last inequality holds since $\tilde{T}_i \le T_i$. By plugging this in we get: $16 \le \exp\left(\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right)$. We now condition 15 on $\{X_i, e_i\}_{i\in\mathcal{I}_{\text{cal}}}$:

$$
\begin{aligned}
&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha+p_\lambda(X_i)-p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
=&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\right.\right.\\
&\left.\left.+\ t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
=&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\right)\right.\\
&\left.\cdot\ \exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
=&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
&\cdot\ \mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)\\
\le&\exp\left(\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right)\cdot\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}},(X_i,e_i)_{i\in\mathcal{I}_{\text{cal}}}\right)
\end{aligned}
$$

By the law of total expectation, we get:

$$
15\le\exp\left(\frac{|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{2}\right)\cdot\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\,\zeta_\lambda(X_i,e_i)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}}\right)\tag{17}
$$

We now condition on $\{X_i\}_{i\in\mathcal{I}_{\text{cal}}}$ and by the sub-gaussianity of $\zeta_\lambda(X_i,e_i)-w_\lambda(X_i)^{-1}$ we get:

$$
\begin{aligned}
&\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)\left(\zeta_\lambda(X_i,e_i)-w_\lambda(X_i)^{-1}\right)\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}},\{X_i\}_{i\in\mathcal{I}_{\text{cal}}}\right)\\
\le&\exp\left(\frac{t^2}{8}\sum_{i\in\mathcal{I}_{\text{cal}}}\hat{w}_\lambda(X_i)^2\left[\alpha-p_\lambda(X_i)\right]^2\right)\\
\le&\exp\left(\frac{t^2\hat{\gamma}_\lambda^2|\mathcal{I}_{\text{cal}}|}{8}\right)
\end{aligned}
$$

We plug the above in 17 and bound it as follows:

$$
17\le\exp\left(\frac{5|\mathcal{I}_{\text{cal}}|t^2\hat{\gamma}_\lambda^2}{8}\right)\cdot\mathbb{E}\left(\exp\left(t\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)}{w_\lambda(X_i)}\left[\alpha-p_\lambda(X_i)\right]\right)\middle|\mathcal{D}_{\text{tr}}\right)\tag{18}
$$

By combining all of the above, and by following the derivations in the proof of (Gui et al., 2024, Theorem 3) we get

$$
18\le\exp\left(\frac{5|\mathcal{I}_{\text{cal}}|t^2}{8}\left(\pi_\lambda^2+\hat{\gamma}_\lambda^2+\tilde{\gamma}(\lambda)^2\right)+|\mathcal{I}_{\text{cal}}|t\left(\mathbb{E}\left[\left|\frac{\hat{w}_\lambda(X)}{w_\lambda(X)}-\pi_\lambda\right|\middle|\mathcal{D}_{\text{tr}}\right]-\pi_\lambda\Delta\right)\right)\tag{19}
$$

By taking

$$
t:=\frac{\frac{4}{5}\left(\Delta\pi_\lambda-\mathbb{E}\left[\left|\frac{\hat{w}_\lambda(X)}{w_\lambda(X)}-\pi_\lambda\right|\middle|\mathcal{D}_{\text{tr}}\right]\right)}{\pi_\lambda^2+\hat{\gamma}_\lambda^2+\tilde{\gamma}(\lambda)^2},
$$

we have

$$19 \leq \exp\left(-\frac{0.4|\mathcal{I}_{\text{cal}}|\left(\pi_\lambda\Delta - \mathbb{E}\left[\left|\frac{\hat{w}_\lambda(X)}{w_\lambda(X)} - \pi_\lambda\right|\middle|\mathcal{D}_{\text{tr}}\right]\right)^2}{\pi_\lambda^2 + \hat{\gamma}_\lambda^2 + \tilde{\gamma}(\lambda)^2}\right) \leq \delta,$$

where the last inequality follows from the definition of Δ. Therefore,

$$1 - \delta \leq \mathbb{P}(\hat{\alpha}(\tau(\alpha+\Delta)+\varepsilon) > \alpha \mid \mathcal{D}_{\text{tr}}) \leq \mathbb{P}(\hat{\tau} < \tau(\alpha+\Delta)+\varepsilon \mid \mathcal{D}_{\text{tr}})$$

By taking $\varepsilon \to 0$, and by the continuity of the probability measure, we obtain $1 - \delta \leq \mathbb{P}(\hat{\tau} < \tau(\alpha+\Delta)+\varepsilon \mid \mathcal{D}_{\text{tr}})$, which concludes the proof. □

C.3 FUSED CALIBRATION WITH APPROXIMATELY ACCURATE QUANTILES

Following (Gui et al., 2024), instead of proving Theorem 3.2 directly, we provide a more general theorem that implies Theorem 3.2. Our proof relies on the proof of (Gui et al., 2024, Theorem 5).

Theorem C.2. *Fix any $\delta, \alpha \in (0,1)$. Under the same conditions of Theorem 3.1, assume further that there exists a constant $r > 0$ such that:*

1. *For P_X-almost all x: $\sup_{\xi\in[\tau(\alpha),\tau(\alpha+r)+\psi]} w_\xi(x) \leq \gamma$ and $\sup_{\xi\in[\tau(\alpha),\tau(\alpha+r)+\psi]} \hat{w}_\xi(x) \leq \hat{\gamma}$ for some constants $\gamma, \hat{\gamma}, \psi \geq 0$*

2. *for any $\eta \in [0,r]$, for any $\beta \in [\alpha, \alpha+r]$, and for P_X-almost all x:*

$$\mathbb{P}(T < q_\beta(X) + \eta \mid X = x) \leq \beta + B\eta,$$
$$\mathbb{P}(T < q_\beta(X) - \eta \mid X = x) \geq \beta - B\eta.$$

   for some family of oracle functions $\{q_\tau(\cdot)\}_{\tau\in[0,1]}$, and some constant $B > 0$.

3. $\sup_{\beta\in[\alpha,\alpha+r]}\sup_{x\in\mathcal{X}}\left\{\max(B,1)|\hat{q}_{\tau(\beta)}(x) - q_\beta(x)| + \hat{\gamma}\gamma\sqrt{\frac{\log(1/\delta)}{0.4|\mathcal{I}_{\text{cal}}|}}\right\} \leq r.$

Then with probability at least $1-\delta$ over the draw of $\mathcal{D}$, the LPB produced by either Algorithm 1 or by Algorithm 3 satisfies that for P_X-almost all x:

$$\mathbb{P}_{(X,T)\sim P}(T \geq \hat{L}(x) \mid \mathcal{D}, X = x)$$
$$\geq 1 - \alpha - \sup_{\beta\in[\alpha,\alpha+r]}\sup_{x\in\mathcal{X}}\left\{2B|\hat{q}_{\tau(\beta)}(x) - q_\beta(x)|\right\} - \hat{\gamma}\gamma\sqrt{\frac{\log(1/\delta)}{0.4|\mathcal{I}_{\text{cal}}|}}.$$

Proof. Similarly to Appendix C.2, the following proof guarantees the validity of both the `focused` and `fused`. For each method, we embrace its appropriate notations as follows. For the `focused` method, we use $\hat{\alpha}(\cdot) := \hat{\alpha}_{\text{focused}}(\cdot)$, $\hat{\tau} := \hat{\tau}_{\text{focused}}$, and $\zeta_\lambda(X_i, e_i) = e_i$. For the `fused` method, we consider $\hat{\alpha}(\cdot) := \hat{\alpha}_{\text{fused}}(\cdot)$, $\hat{\tau} := \hat{\tau}_{\text{fused}}$, and $\zeta_\lambda(X_i, e_i) = \mathbb{I}\{s_\lambda(X_i) = 1 \text{ or } e_i = 1\}$. Notice that in both settings, the weights are formulated as $w_\tau(x) := 1/\mathbb{P}(\zeta_\lambda(X,e) = 1 \mid X = x)$. As in the proof in Appendix C.2, our claims hold for each choice of method and its corresponding terms.

We begin by defining the error terms:

$$\mathcal{E} = \sup_{\beta\in[\alpha,\alpha+r]}\sup_{x\in\mathcal{X}}\left\{|\hat{q}_{\tau(\beta)}(x) - q_\beta(x)|\right\}, \quad \Delta = B\mathcal{E} + \hat{\gamma}\gamma\sqrt{\frac{\log(1/\delta)}{0.4|\mathcal{I}_{\text{cal}}|}}. \tag{20}$$

For simplicity, we also denote $\lambda := \tau(\alpha+\Delta)+\varepsilon$ for some $\varepsilon > 0$. We further adopt the definition of $p_a(x) := \mathbb{P}(T \leq \hat{q}_a(X) \mid X = x)$ from (Gui et al., 2024). The rest of the proof follows from (Gui et al., 2024, Theorem 5), except for substituting $\hat{f}$ by $\hat{q}$, $\mathbb{I}\{C_i \geq \hat{f}_\lambda(X_i)\}$ by $\zeta(X_i, e_i)$, $\mathbb{I}\{C_i \geq \hat{f}_\lambda(X_i) > T_i\}$ by $\mathbb{I}\{\zeta(X_i, e_i) = 1, \hat{f}_\lambda(X_i) > \tilde{T}_i\}$, and $\mathbb{I}\{T_i < \hat{f}_\lambda(X_i)\}$ by $\mathbb{I}\{\tilde{T}_i < \hat{f}_\lambda(X_i)\}$. All derivations from the proof of (Gui et al., 2024, Theorem 5) apply to our setting except for the upper bound of the following term, which we develop next.

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left\{t\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)}{w_\lambda(X_i)}\left(p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right)\middle|\mathcal{D}_{\text{tr}},\{X_i\}_{i\in\mathcal{I}_{\text{cal}}}\right\}\right] \\
&\leq \exp\left\{\frac{t^2}{2}\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)^2}{w_\lambda(X_i)^2}+t\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)}{w_\lambda(X_i)}\mathbb{E}\left[\left(p_\lambda(X_i)-\mathbb{I}\{\tilde{T}_i<\hat{q}_\lambda(X_i)\}\right)\middle|\mathcal{D}_{\text{tr}},\{X_i\}_{i\in\mathcal{I}_{\text{cal}}}\right]\right\} \\
&\leq \exp\left\{\frac{t^2}{2}\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)^2}{w_\lambda(X_i)^2}+t\sum_{i\in\mathcal{I}_{\text{cal}}}\frac{\hat{w}_\lambda(X_i)}{w_\lambda(X_i)}\left[\mathbb{P}(T_i<\hat{q}_\lambda(X_i)|\mathcal{D}_{\text{tr}},X_i)-\mathbb{P}(\tilde{T}_i<\hat{q}_\lambda(X_i)|\mathcal{D}_{\text{tr}},X_i)\right]\right\} \\
&\leq \exp\left\{\frac{t^2}{2}|\mathcal{I}_{\text{cal}}|\hat{\gamma}^2\right\}
\end{aligned}
$$

Above, the first inequality follows from Hoeffding's inequality and the last one holds since $\tilde{T}_i \leq T_i$, and due to the upper bounds of the weights $\hat{w}_\lambda(\cdot)$. We plug in this bound in the derivations of (Gui et al., 2024) and get that

$$
\mathbb{P}(\hat{\alpha}(\tau(\alpha+\Delta)+\varepsilon)>\alpha \mid \mathcal{D}_{\text{tr}}) \leq \exp\left\{-\frac{|\mathcal{I}_{\text{cal}}|t(\Delta-B\mathcal{E})}{\gamma}+\frac{5|\mathcal{I}_{\text{cal}}|\hat{\gamma}^2t^2}{8}\right\}
$$

Therefore, we take $t=\frac{4}{5\gamma\hat{\gamma}^2}(\Delta-B\mathcal{E})$ we get

$$
\mathbb{P}(\hat{\alpha}(\tau(\alpha+\Delta)+\varepsilon)>\alpha \mid \mathcal{D}_{\text{tr}}) \leq \exp\left\{-0.4|\mathcal{I}_{\text{cal}}|\frac{1}{\gamma^2\hat{\gamma}^2}(\Delta-B\mathcal{E})^2\right\} \leq \delta
$$

where the last inequality follows from the definition of Δ from Equation 20. Therefore, we get that with probability $1-\delta$ over the draw of $\mathcal{D}$: $\mathbb{P}(\hat{\alpha}(\tau(\alpha+\Delta)+\varepsilon)>\alpha)$. This implies $\tau(\alpha+\Delta)+\varepsilon>\hat{\tau}$, and thus:

$$
\mathbb{P}(\tau(\alpha+\Delta)+\varepsilon>\hat{\tau}) \geq 1-\delta.
$$

As in (Gui et al., 2024), we take $\varepsilon \to 0$, and by the continuity of the probability measure, we get that $\tau(\alpha+\Delta)+\varepsilon>\hat{\tau}$ holds with probability at least $1-\delta$, which concludes the proof.

□

D Experimental setup

In all experiments, the dataset was split into four parts: 60% for training, 20% for calibration, 10% for validation (used for early stopping), and 10% for testing to evaluate performance. Synthetic data was generated through distribution simulations, as outlined in Table 1, while the process for acquiring the TCGA data is described in Section D.3. The synthetic data generation function and the processed TCGA-BRCA dataset are available in the supplementary code. The TCGA-BRCA dataset was further normalized and imputed using the dataset pre-processing code by Ketenci et al. (2023), utilizing the pre-processing of Nagpal et al. (2021a) for data imputation and normalization.

In all experiments, we approximated the distribution of $T \mid X$ using the *DeepSurv* method (Katzman et al., 2018), implemented in the *pycox* package (Kvamme et al., 2019), implemented using a PyTorch MLP regressor with ReLU activation, early stopping (triggered after 20 epochs without improvement), and a training cycle of 1000 epochs. The Adam optimizer optimized the model with parameters $lr=1e-3$, $\beta_1=0.9$ and $\beta_2=0.999$, a batch size of 256, dropout layers with a rate of $p=0.1$, batch normalization layers, and varying configurations of hidden layers, detailed in Table 2. These configurations were selected to be similar to those found in the PyCox notebooks, with the TCGA-BRCA setting getting a deeper model to account for its more complex and interconnected nature.

Additionally, we employed *scikit-learn* (Pedregosa et al., 2011) to train a Random Forrest Classifiers with a max depths of 4 and 2, to estimate the weights $\hat{w}_\tau$ and the indicator $\hat{s}_\tau$ respectively. The higher max depth for estimating $\hat{w}_\tau$ is meant to make conditional calibration more accurate, promoting validity. In comparison, the lower max depth of $\hat{s}_\tau$ is there to reduce potential overfitting, that might lead to overly conservative estimators.

Table 2: Parameters specifying the number of hidden layers of the MLP regressor employed for the *DeepSurv* survival analysis model in the synthetic and TCGA settings.

Setting	Hidden layers
Synthetic settings	$[32]$
TCGA-BRCA	$[32, 32]$

D.1 Machine specifications

The hardware and OS used for the experiments are as follows.

- **CPU:** AMD EPYC 7443 24-Core Processor
- **GPU:** NVIDIA RTX A6000
- **OS:** Ubuntu 20.04

D.2 Computational efficiency

Note that the values of τ that cause a shift in $\hat{\alpha}(\tau)$ are those for which $\tilde{T}_i = \hat{q}_\tau(X_i)$, for $i \in \mathcal{I}_{\text{cal}}$. As a result, to compute the miscoverage estimator one need only check these values and $\tau = 0$. As such, the computational complexity of the calibration is $O(|\mathcal{I}_{\text{cal}}|)$, and so the brunt of computing is allocated towards the model training, both for the quantile and weight estimation.

D.3 TCGA acquisition and preprocessing details

The TCGA-BRCA cohort is accessible through the GDC portal. To create a dataset suitable for *DeepSurv*, we exclude patients without clinically relevant Estrogen Receptor (ER) status or scanned biopsies. We then merge several tables and select the following clinically significant features for model training, imputing the Progesteron Receptor (PR) value with the ER value, and the rest of the features with the value value -1:

- ER status $\in \{0, 1\}$
- PR status $\in \{0, 1\}$
- Her2 status $\in \{0, 1\}$
- Presence of infiltrating ductal carcinoma $\in \{0, 1\}$
- Presence of infiltrating lobular carcinoma $\in \{0, 1\}$
- PGR gene expression $\in [0, 1]$
- ESR1 gene expression $\in [0, 1]$
- ERBB2 gene expression as defined by Desmedt et al. (2008) $\in [0, 1]$
- Gender $\in \{0, 1\}$
- Age $\in \{26, \cdots, 90\}$
- Tamoxifen drug sensitivity $\in [0, 1]$
- Lapatinib drug sensitivity $\in [0, 1]$

Additionally, we incorporate the cancer grade of each patient by combing through their written medical reports. Finally, we process a single whole-slide H&E biopsy image for each patient, which can vary significantly in size, typically around 300,000 pixels. To identify the tissue regions, we apply Otsu's segmentation method. After segmentation, we use the *GigaPath* foundation model from Xu et al. (2024) to generate compact embedding vectors that capture the essential features of each biopsy. These embeddings are then further reduced to 3 dimensions using PCA (Principal Component Analysis). All of the above features are concatenated to create a covariate vector X of size 19. The survival time T is defined as the time to mortality, with the censored time $\tilde{T}$, and event indicator e given by the dataset.