KSP: Kolmogorov-Smirnov metric-based Post-Hoc Calibration for Survival Analysis

Jeongho Park[1] **Daheen Kim**[1] **Cheoljun Kim**[2]

Hyungbin Park[2] **Sangwook Kang**[1] **Gwangsu Kim**[2,3,*]

[1]Department of Statistics and Data Science, Yonsei University, Seoul, Republic of Korea
[2]Department of Statistics, Jeonbuk National University, Jeonju, Republic of Korea
[3]Research Institute for Materials and Energy Sciences, Jeonbuk National University
[1]{wjdgh4325, ddaccong99, kanggi1}@yonsei.ac.kr
[2]{fefe2, hbag89503, s88012}@jbnu.ac.kr

Abstract

We propose a new calibration method for survival models based on the Kolmogorov–Smirnov (KS) metric. Existing approaches—including conformal prediction, D-calibration, and Kaplan–Meier (KM)-based methods—often rely on heuristic binning or additional nonparametric estimators, which undermine their adaptability to continuous-time settings and complex model outputs. To address these limitations, we introduce a streamlined *KS metric-based post-processing* framework (KSP) that calibrates survival predictions without relying on discretization or KM estimation. This design enhances flexibility and broad applicability. We conduct extensive experiments on diverse real-world datasets using a variety of survival models. Empirical results demonstrate that our method consistently improves calibration performance over existing methods while maintaining high predictive accuracy. We also provide a theoretical analysis of the KS metric and discuss extensions to in-processing settings.

1 Introduction

Calibration plays a vital role in ensuring reliable risk estimation for decision-making in survival analysis. Since poor calibration can misrepresent failure risk, leading to suboptimal or misguided decisions, it is especially critical in high-stakes areas such as healthcare and infrastructure systems. Recently, deep neural networks (DNNs) have significantly improved predictive performance in survival analysis (Wiegrebe et al., 2024), as evidenced by models such as DeepSurv (Kim et al., 2019), Transformer-based architectures (Hu et al., 2021), and SurvTrace (Wang and Sun, 2022). These models are typically evaluated using the concordance index (C-index; Harrell Jr et al., 1996), which measures a model's ability to correctly rank individuals. Although DNNs improve discrimination, they often suffer from over-confidence, a widely recognized issue in classification (Guo et al., 2017; Kumar et al., 2018; Mukhoti et al., 2020), which results in poor calibration. Nevertheless, survival calibration remains underdeveloped and is further complicated by censoring and time-dependent risks.

One notable contribution to survival calibration is D-calibration (Distributional calibration; Haider et al., 2020), which compares the predicted probability with the observed failure proportion within

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

predefined intervals. Building on this idea, X-cal (eXplicit calibration; Goldstein et al., 2020) incorporates D-calibration directly into model training. However, it relies on binning, which may underestimate calibration error (Kumar et al., 2019) and often substantially reduces discrimination performance (Qi et al., 2024a; Park et al., 2025). To reduce such trade-offs, recent studies (Avati et al., 2020; Fuhlert et al., 2022; Qi et al., 2024a,b; Lee et al., 2024) have aimed to improve calibration while maintaining predictive performance. Some leverage conformal inference (Qi et al., 2024a,b) or outcome-aware sampling (Lee et al., 2024), but their reliance on fixed sampling schemes or predefined percentiles may limit adaptability.

A promising direction for addressing these limits comes from classical statistical tools. The Kolmogorov–Smirnov (KS) metric has long been used to assess discrepancies between empirical and estimated distributions (Fernández and Gretton, 2019), and has also been employed as a calibration metric in classification (Gupta et al., 2021; Arrieta-Ibarra et al., 2022), offering a nonparametric, distribution-level evaluation of calibration error. Although the KS metric has also been applied in survival analysis (Rao, 1998; Fleming et al., 1980; Schumacher, 1984; Wu, 2018; Ansin, 2015; Cox and Snell, 1968), such studies have predominantly focused on testing, with limited attention given to its role in assessing or improving calibration error. A recent study (Park et al., 2025) introduced a KS metric for survival models, using the maximum deviation between predicted and empirical survival functions as a calibration measure. Building on this insight, we propose an efficient post-processing method that leverages the KS metric to improve calibration without compromising predictive accuracy. Our approach avoids binning, surrogate losses, and quantile estimation, providing a bin-free, global measure of calibration error.

Our main contributions are as follows:

1. We introduce a simple and scalable post-processing method for calibrating survival models, eliminating the need for surrogate losses, sampling, and quantile estimation.
2. We provide theoretical and algorithmic insights into the KS metric, including its connection to calibration in survival analysis.
3. We conduct extensive empirical evaluations across diverse real-world datasets and models, showing that the method improves calibration without sacrificing predictive performance.

The remainder of this paper reviews the background and motivation for calibration, examines the theoretical properties of the KS metric, and presents the post-processing method along with experimental results. All proofs of the theorem and proposition are deferred to Appendices A and C.

2 Related work

Evaluation In medical research, the calibration of survival models is often assessed graphically using calibration curves, where perfect calibration aligns with the 45-degree line representing equality between observed and predicted probabilities (Alonzo, 2009; Crowson et al., 2016). Goodness-of-fit can be evaluated using a Hosmer–Lemeshow-type test (Hosmer and Lemeshow, 1980). For DNN-based survival models, 1-calibration, which evaluates predicted survival probabilities at a specific time point, is commonly used (Yan et al., 2022; Li et al., 2023; Xia et al., 2023), though it does not capture calibration across the full survival distribution. To address this, D-calibration was introduced to assess calibration over the entire distribution (Haider et al., 2020). However, its reliance on fixed bins may underestimate calibration error (Kumar et al., 2019). Yanagisawa (2023) uses KM-calibration as an evaluation metric.

In-processing Calibration has also been incorporated into the model training process (Goldstein et al., 2020; Chapfuwa et al., 2020; Avati et al., 2020; Lee et al., 2024; Park et al., 2025). X-cal (Goldstein et al., 2020) introduces a bin-based D-calibration penalty term, while S-cal (Park et al., 2025) replaces it with a random-interval-based term. Chapfuwa et al. (2020) integrates KM estimates directly into training. These methods enhance marginal calibration, but often at the cost of reduced discrimination performance. A notable exception is the contrastive learning-based method by Lee et al. (2024), which aims to enhance discrimination without sacrificing calibration.

Post-processing To balance calibration and discrimination, post-processing approaches have been proposed primarily using conformal prediction techniques (Candès et al., 2023; Qi et al., 2024a,b; Gui et al., 2024; Qin et al., 2025; Davidov et al., 2025). The CSD method (Qi et al., 2024a) refines

survival probabilities via KM-based percentile adjustments, ensuring marginal calibration across models. It offers a model-agnostic, post-hoc calibration framework. Qi et al. (2024b) further extends this to calibrate conditional distributions for censored data.

3 Discrimination and calibration in survival analysis

We consider right-censored survival data, where the non-negative failure time T is subject to censoring. Let C be the censoring time, and define the observed time as $Y = \min(T, C)$ and the censoring indicator as $\delta = \mathbb{I}(T \leq C)$, where $\mathbb{I}(A)$ denotes the indicator function, taking the value 1 when A is true and 0 otherwise. Each observation consists of $(Y, \delta, \boldsymbol{z})$, where $\boldsymbol{z}$ is a vector of covariates. Assuming independent censoring given covariates, $T \perp\!\!\!\perp C \mid \boldsymbol{z}$, the full dataset is denoted by $\{(Y_i, \delta_i, \boldsymbol{z}_i)\}_{i=1}^N$ for a sample size of N. Let $F(t \mid \boldsymbol{z}) = \mathbb{P}(T \leq t \mid \boldsymbol{z})$ denote the conditional cumulative distribution function (CDF). The corresponding survival function S, hazard function λ, and cumulative hazard function Λ are as follows:

$$\lambda(t \mid \boldsymbol{z}) = \lim_{\Delta \downarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta \mid T \geq t, \boldsymbol{z})}{\Delta} \text{ and } S(t \mid \boldsymbol{z}) = \mathbb{P}(T > t \mid \boldsymbol{z}) = 1 - F(t \mid \boldsymbol{z}),$$

with the relationship $S(t \mid \boldsymbol{z}) = \exp\{-\Lambda(t \mid \boldsymbol{z})\}$, where $\Lambda(t \mid \boldsymbol{z}) = \int_0^t \lambda(s \mid \boldsymbol{z})\, ds$.

3.1 Discrimination

The C-index is a standard discrimination metric in survival analysis, assessing how well the model ranks individuals by risk. Harrell's C-index (Harrell Jr et al., 1996) is defined as

$$\text{C-index} = \frac{\sum_{i<j} \delta_i \mathbb{I}(Y_i < Y_j)\, \mathbb{I}(M_i < M_j)}{\sum_{i<j} \delta_i \mathbb{I}(Y_i < Y_j)}$$

where M_i denotes the predicted mean or a related quantity. The C-index ranges from 0 to 1, with higher values indicating better discrimination. In our case, we use the predicted mean as M_i.

3.2 Calibration and metrics

Calibration in survival analysis refers to the agreement between the estimated survival distributions and actual event occurrences. While discrimination is important, calibration must also be considered to assess the uncertainty in predicted survival probabilities. A well-calibrated model should accurately estimate the probability of an event occurring by a given time. One fundamental approach to assessing calibration relies on the estimated CDF of event times. Specifically, if $T \sim F$, then $F(T) \sim \text{Unif}[0, 1]$; that is, the CDF evaluated at the true event time follows a uniform distribution over the interval $[0, 1]$. This probabilistic property forms the basis of several calibration metrics, including D-calibration and KM-calibration.

3.2.1 D-calibration

Haider et al. (2020) introduced D-calibration, a measure designed to evaluate calibration under censoring by leveraging the distributional properties of the estimated CDF. For uncensored data, the CDF values satisfy $F(T \mid \boldsymbol{z}) \sim \text{Unif}[0, 1]$. In the presence of right censoring (i.e., $T > C$), the distribution becomes $F(T \mid \boldsymbol{z}) \sim \text{Unif}[F(C \mid \boldsymbol{z}), 1]$. Combining both cases, the following equality holds:

$$\mathbb{E}_{Y,\delta,\boldsymbol{z}}\left[\delta \mathbb{I}(F(Y \mid \boldsymbol{z}) \in I) + (1 - \delta)\mathbb{P}(F(T \mid \boldsymbol{z}) \in I)\right] = |I| \tag{1}$$

Here, $I = [a, b] \subseteq [0, 1]$ is a subinterval. The term $\mathbb{P}(F(T \mid \boldsymbol{z}) \in I)$ is computed as $\mathbb{P}(F(T \mid \boldsymbol{z}) \in I) = \frac{b - F(Y|\boldsymbol{z})}{1 - F(Y|\boldsymbol{z})} \mathbb{I}(F(Y \mid \boldsymbol{z}) \in I) + \frac{b-a}{1-F(Y|\boldsymbol{z})} \mathbb{I}(F(Y \mid \boldsymbol{z}) < a)$. Based on this, D-cal is defined as the sum of bin-wise squared differences:

$$\text{D-cal} = \sum_{I \in \mathcal{I}} \left(\mathbb{E}_{Y,\delta,\boldsymbol{z}}\left[\delta \mathbb{I}\left(\hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z}) \in I\right) + (1 - \delta)\mathbb{P}\left(\hat{F}_{\boldsymbol{\theta}}(T \mid \boldsymbol{z}) \in I\right)\right] - |I| \right)^2$$

where $\mathcal{I} = \{I_1, \ldots, I_B\}$ is a set of $B\,(> 0)$ disjoint intervals partitioning $[0, 1]$, and $\hat{F}_{\boldsymbol{\theta}}$ denotes the estimated CDF with model parameters $\boldsymbol{\theta}$. D-cal measures the discrepancy between the observed proportion of estimated CDF values falling into each bin and the corresponding predicted probability.

3.2.2 KM-calibration

While D-calibration evaluates calibration in the probability space by assessing the uniformity of the estimated CDF values, an alternative approach focuses on the time domain. Specifically, Yanagisawa (2023) and Qi et al. (2024a) proposed calibration metrics based on the KM estimator. One such metric is defined as KM-cal $= \frac{1}{t_{\max}} \int_0^{t_{\max}} \left(S_{\mathrm{KM}}(t) - \mathbb{E}_{\boldsymbol{z}} \left[\hat{S}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}) \right] \right)^2 \mathrm{d}t$, where $S_{\mathrm{KM}}(t)$ is the KM estimator and $\hat{S}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}) = 1 - \hat{F}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z})$ is the model-based survival function. This metric assumes that the KM estimator is well-calibrated and measures the squared discrepancy between the empirical and predicted survival probabilities over time.

4 Motivation for new algorithm

Several recent algorithms leverage D-calibration and KM-calibration metrics, including X-cal (Goldstein et al., 2020), SFM (Survival Function Matching; Chapfuwa et al., 2020), CSD (Conformalized Survival Distributions; Qi et al., 2024a), and CSD-iPOT (individual survival Probability at Observed Time; Qi et al., 2024b). X-cal and SFM incorporate empirical variants of D-cal and KM-cal as penalty terms during model training. For example, X-cal replaces the discontinuous indicator $\mathbb{I}(x \in [a, b])$ with a smooth surrogate $\zeta_\gamma(x; [a, b]) = \frac{1}{1+\exp(-\gamma(x-a)(b-x))}$ (for $\gamma > 0$), which enables gradient-based optimization. In contrast, CSD and CSD-iPOT adjust estimated survival distributions using conformal inference over quantile levels. These approaches are effective in reducing calibration error while maintaining predictive accuracy, but they have some limitations.

X-cal Binning-based methods, such as those underlying D-cal and X-cal, may underestimate calibration error (Kumar et al., 2019), a limitation that extends to the survival setting. Additionally, penalty-based calibration can lead to degradation in discriminative performance, as observed in Qi et al. (2024a); Park et al. (2025).

SFM This method involves computing $\mathbb{E}_{\boldsymbol{z}}[\hat{S}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z})]$, which can be computationally intensive. Moreover, when the model's estimated survival function is already better calibrated than the KM estimator, the KM-based correction may not further reduce calibration error.

CSD and CSD-iPOT CSD relies on the KM estimator for handling censored observations, whereas CSD-iPOT avoids KM-based estimation. However, both methods involve sampling procedures, which can introduce additional variance, especially in the tail regions, and may result in non-monotonic quantiles (i.e., quantile crossing).

These limitations motivate a calibration approach that (i) avoids discretization and binning, (ii) preserves discriminative performance (e.g., C-index), (iii) does not require computationally expensive sampling, and (iv) remains stable in tail regions.

The KS metric is a classical tool for quantifying discrepancies between empirical and target distributions. In the calibration context, it has been applied to classification tasks (Gupta et al., 2021; Arrieta-Ibarra et al., 2022) as a nonparametric, bin-free measure of calibration error. Among existing tools, Cox–Snell residuals (Cox and Snell, 1968) can be used to evaluate calibration error. However, they rely on the cumulative hazard function and are less practical than CDF-based approaches. Therefore, we focus on an alternative form of calibration error. Previous work of Park et al. (2025) proposed using the KS metric as a calibration measure in survival analysis. We refer to this metric as KS-cal and build upon it by introducing a post-processing algorithm that improves calibration while preserving predictive accuracy.

Let $U_i = \hat{F}_{\boldsymbol{\theta}}(Y_i \mid \boldsymbol{z}_i)$. The KS-cal is defined as

$$\text{KS-cal} = \sup_{x \in [0,1]} \left| \tilde{F}(x) - x \right|, \text{ where } \tilde{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(U_i \le x) \left\{ \delta_i + (1 - \delta_i) \frac{x - U_i}{1 - U_i} \right\}. \tag{2}$$

For finite samples, we approximate KS-cal using the evaluation points $q_j = \hat{F}_{\boldsymbol{\theta}}(y_j \mid \boldsymbol{z}_j)$ as KS-cal $= \max_{1 \le j \le N} D_j$, where $D_j = \max\{D_{j,u}, D_{j,l}\}$ with $D_{j,u} = |\tilde{F}(q_j) - q_j|$ and $D_{j,l} = |\tilde{F}(q_j) - \delta_j/N - q_j|$. The term $D_{j,l}$ accounts for the left-limit behavior of $\tilde{F}$ at uncensored points.

We observe that $\tilde{F}(x)$ exhibits jumps at uncensored points while increasing linearly over censored points. Figure 1 illustrates this behavior, showing how D_j is computed at uncensored observations

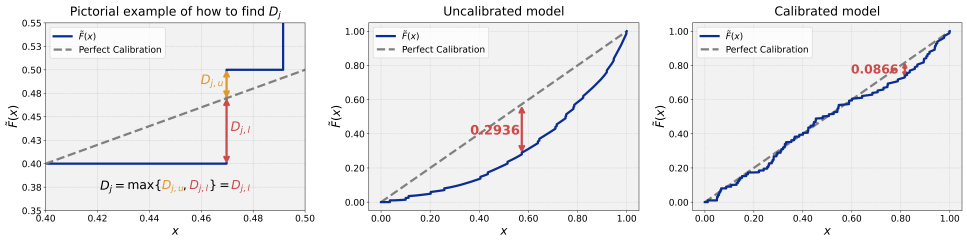


Figure 1: Examples of $\tilde{F}(x)$. The solid blue line represents $\tilde{F}(x)$, and the dashed line shows the identity line $y = x$. The left panel indicates how D_j is computed at uncensored observations. The middle and right panels illustrate uncalibrated and calibrated models, respectively, with respect to KS-cal. Red arrows indicate points of evaluation on KS-cal.

and contrasting uncalibrated and calibrated models in terms of KS-cal. This metric offers a bin-free assessment of calibration error and is conceptually related to D-calibration (Haider et al., 2020), although key differences in how calibration is measured remain. D-cal evaluates deviations within fixed bins, whereas KS-cal captures the maximum discrepancy at data-adaptive points where miscalibration is most evident, as determined by $\hat{F}_{\boldsymbol{\theta}}$. In the following section, we formalize the theoretical properties of the KS-cal, including consistency and convergence behavior.

4.1 Property of KS-cal

We now analyze the theoretical behavior of KS-cal. We begin with the simplified setting without covariates, i.e., $U_i = \hat{F}_{\boldsymbol{\theta}}(Y_i)$, where calibration corresponds to $\hat{F}_{\boldsymbol{\theta}} = F$, with F denoting the true CDF of T. The following result has been established in the previous work of Park et al. (2025).

Proposition 4.1. *Assume the Regularity Conditions in Appendix A. Then calibration holds if and only if*

$$\sup_{x\in[0,1]} |\tilde{F}(x) - x| = o_p(1) \text{ as } N \to \infty.$$

We now extend this result to the setting with covariates, i.e., $U_i = \hat{F}_{\boldsymbol{\theta}}(Y_i \mid \boldsymbol{z}_i)$. Following Goldstein et al. (2020), calibration with covariates is defined as:

$$\mathbb{E}_{Y,\delta}\mathbb{E}_{\boldsymbol{z}}\left[\mathbb{I}\left(\hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z}) \le x\right)\left\{\delta + (1-\delta)\frac{x - \hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z})}{1 - \hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z})}\right\}\right] = x, \ \forall x \in [0,1]. \tag{3}$$

Then, we have the following theorem.

Theorem 4.1. *Assume the Regularity Conditions in Appendix A. Then calibration (with/without covariates) holds if and only if*

$$\sup_{x\in[0,1]} |\tilde{F}(x) - x| = o_p(1) \text{ as } N \to \infty.$$

This result implies that $\tilde{F}$ converges to the CDF of Unif$[0,1]$ when the model is calibrated, where the theorem has two folds (with/without covariates). Therefore, we can establish KS-cal as a theoretically grounded, bin-free metric for evaluating calibration in survival models. Moreover, under the regularity conditions in Appendix A, the convergence of $\tilde{F}$ (and thus KS-cal) occurs at the rate of $O_p(N^{-1/2})$ up to $\log N$.

Remark. Based on Eqn. (3), an alternative calibration metric termed S-cal (Summation-based Calibration) has been proposed (Qi et al., 2024a,b; Park et al., 2025). This metric captures cumulative deviations from ideal calibration and is defined as:

$$\text{S-cal} = \mathbb{E}_s\left[\left(\mathbb{E}_{Y,\delta,\boldsymbol{z}}\left[\mathbb{I}\left(\hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z}) \le s\right)\left\{\delta + (1-\delta)\frac{s - \hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z})}{1 - \hat{F}_{\boldsymbol{\theta}}(Y \mid \boldsymbol{z})}\right\}\right] - s\right)^2\right]$$

S-cal, D-cal, and KS-cal are all variants of D-calibration, derived from the same foundational definition, but they emphasize different aspects of calibration behavior. We include all three in our evaluation to provide a comprehensive view of model calibration.

5 Calibration method based on KS-cal

In this work, we propose a post-hoc calibration method based on the KS-cal introduced in Section 4. We also briefly describe how the KS-cal can be incorporated as a penalty term during model training.

5.1 Post-processing

We propose KS-cal based post-processing (KSP), similar to Platt scaling (Platt, 1999) in classification tasks. We simply transform the original $\hat{F}_{\boldsymbol{\theta}}$ into a modified $\hat{F}^*_{\boldsymbol{\theta}} \subseteq [0,1]$ that minimizes the KS-cal. The procedure is summarized in the following algorithm.

Algorithm. KSP

Input: Estimated CDFs $\hat{F}_{\boldsymbol{\theta}}$, strictly monotone increasing link function $G : [0,1] \to (-\infty, \infty)$
Initialize parameters $a\,(>0)$, b, $\alpha\,(>0)$
Sort $\hat{F}_{\boldsymbol{\theta}}$ for computational efficiency
while KS-cal not improved **do**
    Compute transformed CDF: $\hat{F}^*_{\boldsymbol{\theta}} = \left\{ G^{-1}\big(a \cdot G(\hat{F}_{\boldsymbol{\theta}}) + b\big) \right\}^{\alpha}$
    Compute KS-cal on validation set: $\max_{1 \le j \le N} D^*_j$, where D^*_j denotes D_j evaluated using $\hat{F}^*_{\boldsymbol{\theta}}$
    Update (a, b, α) via gradient descent (ADAM) to minimize the KS-cal
end while
Apply final calibrated transformation to the test set using optimized (a, b, α)
Output: Calibrated CDF $\hat{F}^*_{\boldsymbol{\theta}}$

The function G can be any link function satisfying the two predefined conditions described in the Input, such as the logit function, inverse hyperbolic tangent function, or an inverse CDF. Although a DNN-based transformation is feasible, it is challenging to preserve the ordering of the transformed CDFs consistent with the original ones. In this work, we adopt the logit function as G.

The three hyperparameters a, b, and α control different aspects of the CDF transformation: a influences the tails, b shifts the distribution relative to the target x, and α controls nonlinearity. Their effects are visualized and discussed in the Appendix I. As long as $a > 0$ and $\alpha > 0$, the transformation preserves the ordering of CDF values, thereby maintaining the time-dependent C-index (Antolini et al., 2005), consistent with Qi et al. (2024b). Specifically, if $\hat{F}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}_1) > \hat{F}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}_2)$, then the transformed CDF also satisfies $\hat{F}^*_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}_1) > \hat{F}^*_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}_2)$. While the CDF ordering is preserved, the ordering of expected survival times may not be, as the transformation can nonlinearly distort the survival curve. This is particularly relevant when survival curves intersect. In such cases, monotonicity is maintained at each time point, but the area under the curve may change, which in turn affects the expected value. In practice, such changes are typically small, and overall concordance may even improve. We also formalize the conditions under which the mean ordering is preserved. In particular, if the original survival curves do not cross (preserving the order across all times), the transformation preserves the ordering of expected survival times. Theoretically, these are summarized as:

Proposition 5.1. *Let* $\mathbb{E}\left[T^* \mid \boldsymbol{z}\right] = \int_0^\infty S^*\left(t \mid \boldsymbol{z}\right) \mathrm{d}t$ *denote the expected survival time under the KSP, where* $S^*\left(t \mid \boldsymbol{z}\right) = 1 - F^*\left(t \mid \boldsymbol{z}\right)$. *If the original survival curves* $S\left(t \mid \boldsymbol{z}_1\right)$ *and* $S\left(t \mid \boldsymbol{z}_2\right)$ *do not cross, then the ordering of expected survival times is preserved under the KSP:*

$$\mathbb{E}\left[T \mid \boldsymbol{z}_1\right] > \mathbb{E}\left[T \mid \boldsymbol{z}_2\right] \quad \Longleftrightarrow \quad \mathbb{E}\left[T^* \mid \boldsymbol{z}_1\right] > \mathbb{E}\left[T^* \mid \boldsymbol{z}_2\right].$$

The proof is provided in Appendix C. The non-crossing condition holds for some standard models, including the Cox proportional hazards (PH) model (Cox, 1972) and the Weibull Accelerated Failure Time (AFT) model (Stute, 1993). By Proposition 5.1, the C-index is preserved since we use the predicted mean of survival time for the C-index.

Compared to prior methods such as CSD (Qi et al., 2024a) and CSD-iPOT (Qi et al., 2024b), KSP provides better computational efficiency. CSD requires $O\left(N \cdot |\mathcal{P}| \cdot R\right)$ operations, and CSD-iPOT requires $O\left(N \cdot R\right)$, where N and $|\mathcal{P}|$ are the sample size and the number of quantiles, respectively, and R is the number of samples used for censored data. In contrast, KSP requires only $O\left(B \cdot N + N \log N\right)$, which simplifies to $O\left(B \cdot N\right)$ in practice, where B is the number of optimization iterations. The $N \log N$ term accounts for the initial sorting of CDF values, and each iteration involves linear-time operations. Although the number of iterations may increase when the initial calibration error is large, the overall overhead remains modest and can be controlled through the learning rate. This scalability makes KSP particularly suitable for large datasets.

5.2 In-processing

KSP is introduced as a post-processing method to reduce calibration errors while preserving predictive accuracy. Alternatively, the KS-cal defined in (2) can be incorporated directly into training as a penalty term: $\mathcal{P}_k\left(\boldsymbol{\theta}\right) = \sum_{j=1}^{k} \left(D_{(N-j+1)}\right)^2$, where the subscript denotes the order statistic. Unlike the original KS-cal, which considers only the maximum deviation, this variant aggregates the top-k largest deviations. Using the squared form improves gradient stability, making it more suitable for optimization. Since the deviations are evaluated only at observed points, the indicator function in $\tilde{F}\left(x\right)$ naturally disappears when the CDF values are pre-sorted. In contrast, methods like X-cal require a surrogate function to approximate the indicator for gradient-based learning. We refer to this variant as KS-cal(k). While our primary focus is on post-processing, we also explore this in-processing approach in Appendix J. Although the penalty function is not convex, we observe that training remains stable and converges well in practice. However, using excessively small batch sizes can lead to large approximation errors; thus, we recommend employing relatively large batch sizes for reliable performance.

6 Experiment

We first compare post-processing methods, including CSD (Qi et al., 2024a), CSD-iPOT (Qi et al., 2024b), and KSP, across various models and datasets. We also evaluate in-processing approaches such as X-cal (Goldstein et al., 2020) and SFM (Chapfuwa et al., 2020). While SFM originally combines a calibration penalty with a discriminative loss, we use only the calibration component for consistency. For all in-processing methods, we tune the regularization parameter over $\{1, 10, 100, 1000\}$. Experiments are conducted with an Intel Xeon Silver 6226R CPU and an NVIDIA GeForce RTX 3090 GPU.

6.1 Experimental setup

Baselines We consider six baseline models: DeepSurv-based Cox PH model (Katzman et al., 2018), Multi-task Logistic Regression (MTLR; Yu et al., 2011), a parametric model (Goldstein et al., 2020), Survival CRPS (Avati et al., 2020), DeepHit (Lee et al., 2018), and the Weibull AFT model (Stute, 1993). As a reference, we include the KM estimator fitted on the training set and evaluated on the test set, following Qi et al. (2024a,b). The KM estimator serves as an empirical lower bound for calibration error (see Appendix B in Qi et al. (2024a)). Further details on the baseline models are provided in Appendix B.

Datasets We evaluate all methods on ten benchmark datasets: WHAS, METABRIC, GBSG, NACD, NB-SEQ, SUPPORT, MIMIC-III, SEER-liver, SEER-stomach, and SEER-lung. These are grouped by sample size into three categories: **Small** (WHAS, METABRIC, GBSG, NACD), **Medium** (NB-SEQ, SUPPORT, MIMIC-III), and **Large** (SEER-liver, SEER-stomach, SEER-lung). Details on the datasets and preprocessing steps are provided in Appendix D. Each dataset is randomly split into training, validation, and test sets in a 3:1:1 ratio, with balanced censoring rates. All experiments are repeated 30 times with different random seeds. For CSD and CSD-iPOT, we use the validation set as the conformal set to enable a fair comparison with KSP.

Evaluation metrics For predictive accuracy, we report the (time-independent) C-index (Harrell Jr et al., 1996) rather than the time-dependent variant, as the latter may favor our method and CSD-iPOT (Qi et al., 2024b), potentially inflating their discrimination performance. For calibration, we evaluate five metrics: S-cal(20), D-cal(20), KS-cal, KM-cal, and the Integrated Brier Score (IBS;

Table 1: Summary of pairwise comparisons between post-processing methods. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
KSP	20 (12)	**46 (45)**	**46 (43)**	**47 (45)**	**47 (35)**	**44 (25)**
Non-calibrated	18 (1)	13 (7)	14 (6)	13 (5)	13 (10)	16 (1)
Ties	**22**	1	0	0	0	0
KSP	13 (2)	**36 (29)**	**48 (45)**	**51 (42)**	**37 (32)**	**42 (25)**
CSD	**34 (2)**	24 (19)	12 (10)	9 (8)	23 (19)	18 (10)
Ties	13	0	0	0	0	0
KSP	21 (0)	**32 (21)**	**46 (39)**	**44 (29)**	**45 (36)**	**38 (9)**
CSD-iPOT	**25 (1)**	28 (19)	14 (13)	16 (11)	15 (10)	22 (10)
Ties	14	0	0	0	0	0

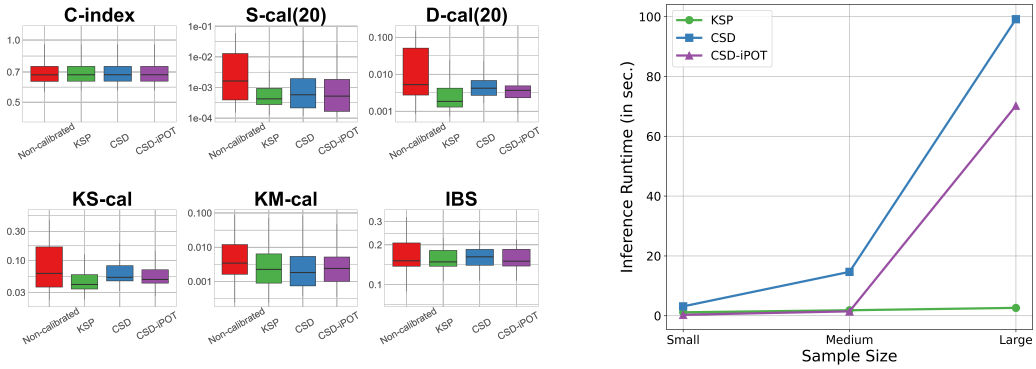


Figure 2: Boxplots of metric values (left) and inference runtime by sample size (right), aggregated across all datasets and models.

Graf et al., 1999). S-cal(20) is computed over 20 equally spaced points $s \in \{0.05, 0.10, ..., 1.00\}$, while D-cal(20) uses 20 equal-width bins. The results based on a 10-bin evaluation are provided in Appendix H. KS-cal is a bin-free metric. KM-cal and IBS assess 1-calibration and offer complementary perspectives. Each model is thus evaluated using one discrimination metric and five calibration metrics. For CSD and CSD-iPOT, we follow the original setup by using $\mathcal{P} = \{0.1, 0.2, ..., 0.9\}$ for quantile adjustment and set $R = 1000$ to sample censored observations.

6.2 Experimental results

Comparison with post-processing methods We conduct $6 \times 10 = 60$ experiments across models and datasets. For each setting, we perform pairwise comparisons among the non-calibrated baseline, CSD, CSD-iPOT, and KSP, counting how often each method outperforms the others. As summarized in Table 1, KSP outperforms the alternatives in approximately 70% of cases. Compared to the non-calibrated baseline, KSP consistently achieves better calibration and significantly improves the C-index in 10 cases, especially for the parametric model and CRPS. Against CSD, KSP shows better calibration, while CSD slightly leads in the C-index due to its guaranty of preserving the time-independent C-index. Compared to CSD-iPOT, KSP offers a similar C-index with clearly better calibration.

KSP's improvements are more evident in D-cal(20) and KS-cal, although it still performs competitively in S-cal(20), which tends to favor CSD and CSD-iPOT due to their quantile-based structure. A similar pattern appears in discretized models like MTLR and DeepHit, where KSP shows slightly weaker calibration, possibly due to structural mismatch. Nonetheless, KSP outperforms in KM-cal and IBS, confirming its overall calibration strength. Figure 3 further illustrates that CSD and CSD-iPOT often exhibit calibration mismatches near the left tail due to boundary interpolation, which D-cal(20) and KS-cal capture effectively. Full results are presented in the Appendix F.

KSP also offers efficiency advantages. As shown in Figure 2, runtime increases with sample size for all methods, but the growth is slower for KSP. This is because CSD and CSD-iPOT require additional bootstrap sampling for the monotonicity of quantiles, while KSP depends only on the number of optimization steps. The complete runtime summary is provided in Appendix K. Figure 2 also presents the distribution of each metric across all experiments, with the median values in the boxplots serving as a key point of comparison. All methods show comparable C-index values, confirming that post-processing can improve calibration without compromising discrimination. KSP consistently achieves the best performance in S-cal(20), D-cal(20), and KS-cal. Although CSD records the lowest median KM-cal, KSP surpasses it more frequently in pairwise comparisons, indicating greater robustness. KSP shares a similar property with CSD-iPOT in preserving time-dependent discrimination, as reflected in their comparable performance on KM-cal and IBS.

Calibration plot Figure 3 shows calibration plots for MTLR and CRPS on the MIMIC-III dataset. Although all three post-processing methods achieve similar quantitative scores, CSD and CSD-iPOT exhibit slight mismatches near the left tail, likely due to boundary interpolation in their quantile adjustments. In contrast, KSP reduces the maximum deviation, resulting in more uniform calibration across the entire range. This pattern aligns with the D-cal(20) and KS-cal results in Table 1, where KSP outperforms other methods by better capturing localized discrepancies. Full results are provided in the Appendix G.

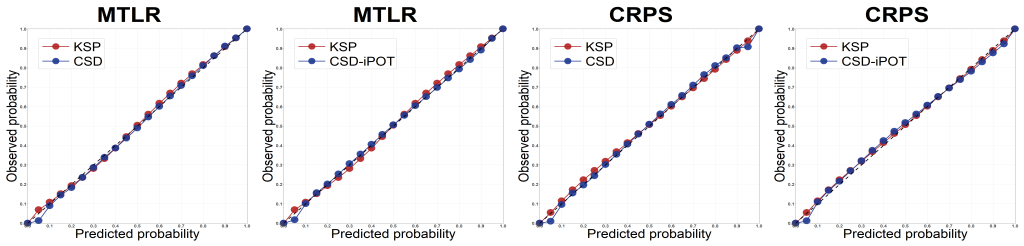


Figure 3: Calibration plot comparison among KSP, CSD, and CSD-iPOT for MTLR and CRPS on the MIMIC-III dataset.

Comparison with in-processing methods We report comparisons with in-processing methods (SFM and X-cal) for CRPS on the MIMIC-III dataset, which illustrate typical trade-off patterns. As λ increases, both in-processing methods reduce calibration error but degrade the C-index. SFM, which penalizes KM-calibration, lowers KM-cal and IBS but performs worse on D-calibration. X-cal improves D-calibration but incurs a greater C-index loss and increases KM-cal and IBS, particularly at $\lambda = 1000$. In contrast, KSP achieves a better trade-off with stable performance across metrics and no need for tuning. While in-processing may be helpful when calibration is the sole priority, KSP offers a more practical and balanced solution when both discrimination and calibration are important. Figure 4 summarizes these results, showing the mean and standard error across multiple λ values.

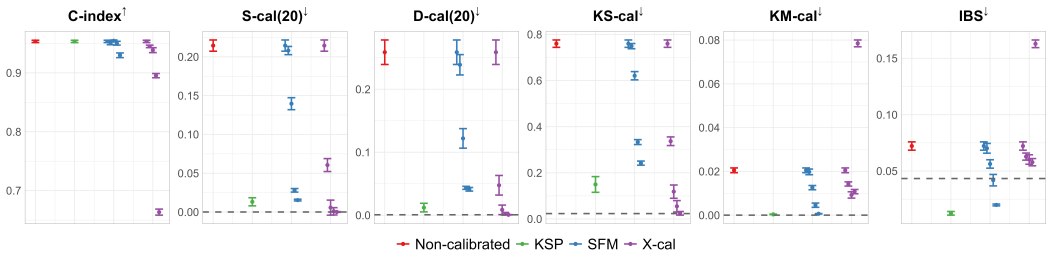


Figure 4: Comparison of KSP and in-processing methods (SFM, X-cal) on the MIMIC-III dataset with CRPS. Points show means and standard errors across $\lambda \in \{0, 1, 10, 100, 1000\}$ for in-processing methods. Dashed lines indicate KM-level calibration errors.

Ablation study We conduct an ablation study to assess the effects of various design choices in the KSP optimization procedure. Specifically, we investigate (1) the choice of link function G, (2) the role of hyperparameters a, b, and α, and (3) alternative loss formulations beyond the default $\max_j D_j^*$,

Table 2: Ablation results on the MIMIC-III dataset using DeepSurv, showing the effect of including hyperparameters a, b, and α in the KSP.

Hyperparameter setting	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
Non-calibrated	0.95839	0.036649	0.064547	0.250907	0.104279	0.276509
b	0.95826	0.011849	0.041974	0.183154	0.077582	0.230868
a, b	0.95826	0.005908	0.025273	0.127957	0.076690	0.213331
a, b, α	0.95826	**0.003305**	**0.007044**	**0.091109**	**0.071617**	**0.203039**

as detailed in the Appendix I. In the main text, we present partial results for (2) on the MIMIC-III dataset using DeepSurv, summarized in Table 2, with full results deferred to the Appendix I. We find that using b alone substantially improves calibration, reducing D-cal(20) and KS-cal by 35% and 27%, respectively, without degrading the C-index. Introducing a and α yields further gains, especially in D-cal(20), by enabling scale adjustment and smoothing. While the magnitude of improvement may vary across datasets and models, all three hyperparameters contribute positively to the KSP procedure. Note that "Non-calibrated" in Table 2 refers to the default setting with $a = 1, b = 0, \alpha = 1$, that is, when KSP is not applied.

7 Conclusion

Our work highlights the effectiveness of KS-cal as a calibration metric and introduces KSP as a practical post-processing method for improving the reliability of survival models. Beyond its empirical and computational advantages, KSP offers a theoretically grounded framework and avoids the limitations of bin-based or sampling-dependent approaches. While KSP shows clear advantages, it tends to be less effective in datasets with heavy ties or severe skewness; however, it is more robust when calibration errors are large and in capturing local discrepancies, such as tails. As an important direction for future research, we aim to extend KSP toward conditional calibration, enabling personalized and subgroup-level reliability assessments. Such developments are particularly valuable for high-stakes domains where accurate uncertainty quantification is critical.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2023-00218377], Global - Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the NRF grant funded by the Ministry of Education [RS-2024–00443714].

References

Alonzo, T. A. (2009). Clinical prediction models: A practical approach to development, validation, and updating: By ewout w. steyerberg. *American Journal of Epidemiology*, 170(4):528.

Ansin, E. (2015). An evaluation of the cox–snell residuals. Master's thesis, Uppsala University, Department of Statistics, Uppsala, Sweden.

Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944.

Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., and Xu, C. (2022). Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23:1–54.

Avati, A., Duan, T., Zhou, S., Jung, K., Shah, N. H., and Ng, A. Y. (2020). Countdown regression: sharp and calibrated survival predictions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 145–155. PMLR.

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

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., and Henao, R. (2020). Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1666–1680.

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

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

Crowson, C. S., Atkinson, E. J., and Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706.

Davidov, H., Feldman, S., Shamai, G., Kimmel, R., and Romano, Y. (2025). Conformalized survival analysis for general right-censored data. In *the International Conference on Learning Representations*. `https://openreview.net/forum?id=JQtuCumAFD`.

Fernández, T. and Gretton, A. (2019). A maximum-mean-discrepancy goodness-of-fit test for censored data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2966–2975. PMLR.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified kolmogorov–smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, pages 607–625.

Fuhlert, P., Ernst, A., Dietrich, E., Westhaeusser, F., Kloiber, K., and Bonn, S. (2022). Deep learning-based discrete calibrated survival prediction. In *Proceedings of the IEEE International Conference on Digital Health*, pages 169–174.

Goldstein, M., Han, X., Puli, A., Perotte, A., and Ranganath, R. (2020). X-cal: explicit calibration for survival analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 18296–18307.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Gui, Y., Hore, R., Ren, Z., and Barber, R. F. (2024). Conformalized survival analysis with adaptive cut-offs. *Biometrika*, 111(2):459–477.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. (2021). Calibration of neural networks using splines. In *the International Conference on Learning Representations*. `https://openreview.net/forum?id=eQe8DEWNN2W`.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. (2020). Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(1):3289–3351.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069.

Hu, S., Fridgeirsson, E., van Wingen, G., and Welling, M. (2021). Transformer-based deep survival analysis. In *Proceedings of Survival Prediction: Algorithms, Challenges and Applications*, pages 132–148.

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., and Roelofs, B. (2021). Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 29768–29779.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24.

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I.-H., and Kim, H. J. (2019). Deep learning-based survival prediction of oral cancer patients. *Scientific Reports*, 9(1):6994.

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, pages 3792–3803.

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the International Conference on Machine Learning*, volume 80, pages 2805–2814. PMLR.

Lee, C., Zame, W., Yoon, J., and Van Der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2314–2321.

Lee, D., Park, H., and Lee, C. (2024). Toward a well-calibrated discrimination via survival outcome-aware contrastive learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 30985–31014.

Li, R., Qu, W., Liu, Q., Tan, Y., Zhang, W., Hao, Y., Jiang, N., Mao, Z., Ye, J., Jiao, J., et al. (2023). Development and validation of a deep learning survival model for cervical adenocarcinoma patients. *BMC Bioinformatics*, 24(1):1–15.

Lin, D. Y. (2007). On the breslow estimator. *Lifetime Data Analysis*, 13(4):471–480.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. (2020). Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299.

Park, J., Kang, S., and Kim, G. (2025). Stochastic explicit calibration algorithm for survival models. *IEEE Access*, 13:78679–78706.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Qi, S.-A., Yu, Y., and Greiner, R. (2024a). Conformalized survival distributions: A generic post-process to increase calibration. In *Proceedings of the International Conference on Machine Learning*, volume 235, pages 41303–41339. PMLR.

Qi, S.-A., Yu, Y., and Greiner, R. (2024b). Toward conditional distribution calibration in survival prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 86180–86225.

Qin, J., Piao, J., Ning, J., and Shen, Y. (2025). Conformal predictive intervals in survival analysis: a resampling approach. *Biometrics*, 81(2):ujaf063.

Rao, M. B. (1998). Survival analysis, techniques for censored and truncated data. *Technometrics*, 40(2):159–160.

Schumacher, M. (1984). Two-sample tests of cramér–von mises- and kolmogorov–smirnov-type for randomly censored data. *International Statistical Review*, 52(3):263–281.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1):89–103.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Wang, Z. and Sun, J. (2022). Survtrace: Transformers for survival analysis with competing events. In *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology & Health Informatics*, pages 1–9.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., and Bender, A. (2024). Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65.

Wu, T. (2018). Randomized survival probability residual for assessing parametric survival models. Master's thesis, University of Saskatchewan, School of Public Health, Saskatoon, Canada.

Xia, Y., Zhang, B., and Zhang, Y. (2023). Deep survival analysis using pseudo values and its application to predict the recurrence of stage iv colorectal cancer after tumor resection. *Computer Methods in Biomechanics and Biomedical Engineering*, 27(15):1–10.

Yan, L., Gao, N., Ai, F., Zhao, Y., Kang, Y., Chen, J., and Weng, Y. (2022). Deep learning models for predicting the survival of patients with chondrosarcoma based on a surveillance, epidemiology, and end results analysis. *Frontiers in Oncology*, 12:967758.

Yanagisawa, H. (2023). Proper scoring rules for survival analysis. In *Proceedings of the International Conference on Machine Learning*, volume 202, pages 39165–39182. PMLR.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We included our limitations on Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the theoretical result in Sections 4.1 and 5.1 and the proofs in Appendices A and C.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental settings, hyperparameters, and evaluation protocols are fully described in the main paper and Appendix. The corresponding code is also publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used are public benchmarks, and we provide open-source code with instructions to reproduce all experiments at `https://github.com/wjdgh4325/KS-cal`, as also noted in the Appendix F.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and Non-calibrateds. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify details on data split in Appendix D, other details in Appendix K.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 1 shows statistical significance, and we provide error bars for experimental results in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We specify compute resources in Section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We follow NeurIPS code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss broader impact in Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk models or datasets are used or released in this work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used publicly available code assets under the MIT License:

    - `https://github.com/rajesh-lab/X-CAL`
    - `https://github.com/shi-ang/CSD`
    - `https://github.com/shi-ang/MakeSurvivalCalibratedAgain`

    All licenses and attributions have been properly acknowledged, and relevant papers are cited in the Appendix F.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We will open new assets after the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: We did not use the LLM for the core parts of the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Broader impacts

Improving survival calibration has broader impacts on clinical decision-making, fairness, resource allocation, and policy development. By producing accurate risk estimates, it supports more reliable and balanced decisions across medical and public health domains.

Code availability

The source code for reproducing our experiments is available at: `https://github.com/wjdgh4325/KS-cal`

A Proofs of Proposition 4.1 and Theorem 4.1

We denote $\hat{F}_{\boldsymbol{\theta}} = F_{\hat{\boldsymbol{\theta}}}$ for notational clarity.

Regularity Conditions

1. With or without covariates, $F_{\hat{\boldsymbol{\theta}}}^{-1}(x)$ exists due to the required properties of $F_{\hat{\boldsymbol{\theta}}}$, such as left-continuity, the existence of a right-hand limit, and monotonicity. For simplicity, we omit the dependence on $\boldsymbol{z}$ in the notation when covariates are present.
2. For the case without covariates, when non-calibration holds,
   - there exists an $x^* \in (0,1)$ such that

$$\eta = \left| F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^* \right| \geq \int_0^{x^*} \left| F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s \right| dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right),$$

   - and for the above x^*, $\displaystyle\int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) < 1$

   where F denotes the true CDF of the failure time, and G denotes the CDF of the censoring time.

Proof of Sufficient Condition in Proposition 4.1

For the proof, we let $\tilde{F}(x) = \frac{1}{N}\sum_{i=1}^{N} \tilde{F}_i(x)$ where $\tilde{F}_i(x) = \mathbb{I}(U_i \leq x)\left\{\delta_i + (1-\delta_i)\frac{x-U_i}{1-U_i}\right\}$ for $0 \leq x \leq 1$, $U_i = F_{\hat{\boldsymbol{\theta}}}(Y_i)$. By Eqn. (8) in Lemma A.2, we have $\mathbb{E}\left[\tilde{F}_i(x)\right] = x$, $\mathbb{E}\left[\tilde{F}_i(x)^2\right] = x - (1-x)\int_0^x \frac{x-s}{1-s} dG\left(F^{-1}(s)\right) \leq 2$ and $0 \leq \tilde{F}_i \leq 1$. Then, by the Bernstein inequality (Van der Vaart, 2000) stating that for an independent random variable X_i such that $\mathbb{E}[X_i] = 0, |X_i| \leq M$,

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{\sum_{i=1}^{N}\mathbb{E}[X_i^2] + M\epsilon/3}\right).$$

For any $0 < \epsilon < 1/2$, we have a set $\mathcal{S}$ of points between $[0,1]$ such that $\{0, \epsilon, 2\epsilon, [1/\epsilon]\epsilon, 1\}$. It implies $\forall t \in [0,1],\ \exists k \in \mathcal{B}$ s.t. $|t-k| \leq \epsilon$. This implies that

$$\left\{\sup_{x\in\mathcal{S}} |\tilde{F}(x) - x| \leq \epsilon\right\} \subset \left\{\sup_{x\in[0,1]} |\tilde{F}(x) - x| \leq 2\epsilon\right\}.$$

Then,

$$\mathbb{P}\left(\sup_{x\in[0,1]} |\tilde{F}(x) - x| > 2\epsilon\right) \leq \mathbb{P}\left(\sup_{x\in\mathcal{S}} |\tilde{F}(x) - x| > \epsilon\right) \leq \sum_{x\in\mathcal{S}} \mathbb{P}\left(|\tilde{F}(x) - x| > \epsilon\right).$$

By the Bernstein inequality, $\mathbb{P}\left(N|\tilde{F}(x)-x|>\epsilon\right) \leq 2\exp(-\epsilon^2/(2N+\epsilon/3))$, due to the symmetry of the absolute value and the expectation of the square is less than 2. This implies

$$\mathbb{P}\left(|\tilde{F}(x)-x|>\epsilon\right) \leq 2\exp\left(-N^2\epsilon^2/(2N+N\epsilon/3)\right) \leq 2\exp\left(-N\epsilon^2/3\right).$$

Finally, we have

$$\mathbb{P}\left(\sup_{x\in[0,1]}|\tilde{F}(x)-x|>2\epsilon\right) \leq \frac{4}{\epsilon}\exp\left(-\frac{N\epsilon^2}{3}\right) \to 0 \tag{4}$$

as $N\to\infty$. □

Proof of Necessary Condition in Proposition 4.1

Assuming the regularity conditions, the fact that non-calibration implies that

$$\tilde{F}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(U_i\leq x)\left\{\delta_i+(1-\delta_i)\frac{x-U_i}{1-U_i}\right\}$$

has the deviation $c\eta>0$ from some x^* is sufficient for the proof. By Lemmas A.2-A.3 and the Regularity Conditions imply that when the non-calibration holds, $\sup_{x\in[0,1]}|\tilde{F}(x)-x| \geq c\eta/2$ with probability approaching 1 as $N\to\infty$. This completes the proof. □

Proof of Theorem 4.1

For simplicity, we let

$$\boldsymbol{z}_i^x(U_i,\delta_i) = \mathbb{I}(U_i\leq x)\left\{\delta_i+(1-\delta_i)\frac{x-U_i}{1-U_i}\right\}.$$

Lemma A.1 and the definition of calibration such as $\mathbb{E}[\boldsymbol{z}_i^x(U_i,\delta_i)]=x$ imply

$$\sup_{x\in[0,1]}|\tilde{F}(x)-x|\to 0$$

in probability as $N\to\infty$. Furthermore, when non-calibration holds such as $\sup_{x\in[0,1]}|\mathbb{E}[\tilde{F}(x)]-x|>\delta$, we have $\mathbb{P}\left(\sup_{x\in[0,1]}|\tilde{F}(x)-x|>\delta/2\right)\to 1$ as $N\to\infty$. This implies that calibration holds if $\mathbb{P}\left(\sup_{x\in[0,1]}|\tilde{F}(x)-x|>\delta/2\right)\to 0$ as $N\to\infty$. It is by contraposition.

Lemma A.1. *Assuming that the Regularity Conditions hold, we have that for any* $0<\epsilon<1/2$

$$\mathbb{P}\left(\sup_{x\in[0,1]}\left|\frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{z}_i^x(U_i,\delta_i)-\mathbb{E}[\boldsymbol{z}_i^x(U_i,\delta_i)]]\right|>\epsilon\right)$$

goes to 0 as $N\to\infty$ *where* $\boldsymbol{z}_i^x(U_i,\delta_i)=\mathbb{I}(U_i\leq x)\left\{\delta_i+(1-\delta_i)\frac{x-U_i}{1-U_i}\right\}$ *and* $U_i=F_{\hat{\boldsymbol{\theta}}}(Y_i\mid \boldsymbol{z}_i)$.

Note that we have $|\boldsymbol{z}_i^x(U_i,\delta_i)|\leq 1$ and $(\boldsymbol{z}_i^x(U_i,\delta_i)-\mathbb{E}[\boldsymbol{z}_i^x(U_i,\delta_i])^2\leq 2$. Then, by the Bernstein inequality (Van der Vaart, 2000) stating that for an independent random variable X_i such that $\mathbb{E}[X_i]=0, |X_i|\leq M$,

$$\mathbb{P}\left(\sum_{i=1}^{N}X_i>\epsilon\right)\leq\exp\left(-\frac{\epsilon^2}{\sum_{i=1}^{N}\mathbb{E}[X_i^2]+M\epsilon/3}\right).$$

For any $0 < \epsilon < 1/2$, we have a set $\mathcal{S}$ of points between $[0,1]$ such that $\{0, \epsilon, 2\epsilon, [1/\epsilon]\epsilon, 1\}$. It implies $\forall t \in [0,1],\ \exists k \in \mathcal{S}$ s.t. $|t-k| \leq \epsilon$. This implies that

$$\begin{aligned}
&\left\{ \sup_{x \in \mathcal{S}} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| \leq \epsilon \right\} \\
&\subset \left\{ \sup_{x \in [0,1]} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| \leq 2\epsilon \right\}.
\end{aligned}$$

Then,

$$\begin{aligned}
& \mathbb{P} \left(\sup_{x \in [0,1]} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > 2\epsilon \right) \\
\leq \ & \mathbb{P} \left(\sup_{x \in \mathcal{S}} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > \epsilon \right) \\
\leq \ & \sum_{x \in \mathcal{S}} \mathbb{P} \left(\left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > \epsilon \right).
\end{aligned}$$

By the symmetric applications of Bernstein inequality,

$$\mathbb{P} \left(N \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > \epsilon \right) \leq 2 \exp \left(-\epsilon^2 / (2N + \epsilon/3) \right),$$

which implies

$$\mathbb{P} \left(\left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > \epsilon \right) \leq 2 \exp \left(-N^2 \epsilon^2 / (2N + N\epsilon/3) \right).$$

This is due to the fact that the square-expectation is less than 2. Finally, we have

$$\mathbb{P} \left(\sup_{x} \left| \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) - \mathbb{E} \left[\boldsymbol{z}_i^x \left(U_i, \delta_i \right) \right] \right] \right| > 2\epsilon \right) \ll \frac{4}{\epsilon} \exp \left(-\frac{N\epsilon^2}{3} \right) \to 0 \tag{5}$$

as $N \to \infty$.

Remark. If we let $\epsilon \equiv \epsilon_N = \frac{M_N \log N}{\sqrt{N}}$ for any diverging sequence M_N, then the probability in the right term of Eqn. (5) approaches 0 as $N \to \infty$. This ensures the convergence rate of $O_p(1/\sqrt{N})$ up to $\log N$.

Lemmas for the proof of Proposition 4.1

Lemma A.2. *Under the Regularity Conditions and $F_{\hat{\boldsymbol{\theta}}} = F$, we have*

$$\begin{aligned}
\mathbb{E} \left[\tilde{F}_i(x) \right] &= x \\
\textit{and } \mathbb{E} \left[\tilde{F}_i(x)^2 \right] &= x - (1-x) \int_0^x \frac{x-s}{1-s} dG \left(F^{-1}(s) \right).
\end{aligned}$$

Proof. Regularity Conditions and calibration of $F_{\hat{\boldsymbol{\theta}}}$ imply that

$$F \left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s) \right) = s.$$

1. Proof of $\mathbb{E} \left[\tilde{F}_i \left(x \right) \right] = x$

Let $T \sim F$, $C \sim G$ and $Z^x(U,\delta) = \mathbb{I}(U \le x)\left\{\delta + (1-\delta)\frac{x-U}{1-U}\right\}$.

Then $\tilde{F}_i(x) \stackrel{\text{def}}{=} Z^x(U_i, \delta_i)$. We have

$$\begin{aligned}
\mathbb{P}(U \le x, \delta = 1) &= \mathbb{P}\left(\min\left(F_{\hat{\boldsymbol{\theta}}}(T), F_{\hat{\boldsymbol{\theta}}}(C)\right) \le x, T \le C\right) = \mathbb{P}\left(F_{\hat{\boldsymbol{\theta}}}(T) \le x, T \le C\right) \\
&= \mathbb{P}\left(T \le F^{-1}_{\hat{\boldsymbol{\theta}}}(x), T \le C\right) = \int_0^{F^{-1}_{\hat{\boldsymbol{\theta}}}(x)} \mathbb{P}(s \le C)\, dF(s) \\
&= \int_0^{F^{-1}_{\hat{\boldsymbol{\theta}}}(x)} [1 - G(s)]\, dF(s) \\
&= \int_0^{x} \left[1 - G\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] dF\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right). \qquad (6)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{P}(U \le x, \delta = 0) &= \mathbb{P}\left(\min\left(F_{\hat{\boldsymbol{\theta}}}(T), F_{\hat{\boldsymbol{\theta}}}(C)\right) \le x, T > C\right) = \mathbb{P}\left(F_{\hat{\boldsymbol{\theta}}}(C) \le x, T > C\right) \\
&= \mathbb{P}\left(C \le F^{-1}_{\hat{\boldsymbol{\theta}}}(x), T > C\right) = \int_0^{F^{-1}_{\hat{\boldsymbol{\theta}}}(x)} \mathbb{P}(T > s)\, dG(s) \\
&= \int_0^{F^{-1}_{\hat{\boldsymbol{\theta}}}(x)} [1 - F(s)]\, dG(s) \\
&= \int_0^{x} \left[1 - F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] dG\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right). \qquad (7)
\end{aligned}$$

By combining (6) and (7), and using integration by parts with respect to $dG(\cdot)$,

$$\begin{aligned}
\mathbb{E}\left[Z^x(U,\delta)\right] &= \int_0^{x} \left[1 - G\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] dF\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \\
&\quad + \int_0^{x} \left[1 - F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] \frac{x-s}{1-s} dG\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \\
&= F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(x)\right) - G\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(x)\right) F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(x)\right) \\
&\quad + \int_0^{x} F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) dG\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \\
&\quad + \int_0^{x} \left[1 - F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] \frac{x-s}{1-s} dG\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \\
&= F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(x)\right) + \int_0^{x} \left\{ - F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(x)\right) + F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \right. \\
&\quad \left. + \left[1 - F\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right)\right] \frac{x-s}{1-s} \right\} dG\left(F^{-1}_{\hat{\boldsymbol{\theta}}}(s)\right) \\
&= x
\end{aligned}$$

2. Proof of $\mathbb{E}\left[\tilde{F}_i(x)^2\right] = x - (1-x)\int_0^{x} \frac{x-s}{1-s} dG\left(F^{-1}(s)\right)$

$$
\begin{aligned}
& \mathbb{E}\left[Z^x(U,\delta)^2\right] \\
= \;& \mathbb{E}\left[\left\{\delta + (1-\delta)\times\frac{x-U}{1-U}\right\}^2 \mathbb{I}\left(U\le x\right)\right] \\
= \;& \mathbb{E}\left[\left\{\delta + (1-\delta)\times\left(\frac{x-U}{1-U}\right)^2\right\} \mathbb{I}\left(U\le x\right)\right] \\
= \;& \int_0^x \left[1-G\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right] dF\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
& + \int_0^x \left[1-F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right] \times \left(\frac{x-s}{1-s}\right)^2 dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
= \;& F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) - F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) G\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) + \int_0^x F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
& + \int_0^x \left(\frac{x-s}{1-s}\right)^2 \left[1-F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right] dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right).
\end{aligned}
$$

It implies that

$$
\begin{aligned}
& \mathbb{E}\left[Z^x(U,\delta)^2\right] \\
= \;& x - xG\left(F^{-1}(x)\right) + \int_0^x s dG\left(F^{-1}(s)\right) + \int_0^x (1-s)\frac{(x-s)^2}{(1-s)^2} dG\left(F^{-1}(s)\right) \\
= \;& x + \int_0^x \left[-x+s+\frac{(x-s)^2}{1-s}\right] dG\left(F^{-1}(s)\right) \\
= \;& x - (1-x)\int_0^x \frac{x-s}{1-s} dG\left(F^{-1}(s)\right). \qquad (8)
\end{aligned}
$$

□

Lemma A.3. *Under the non-calibration assumption such as the Regularity Conditions, we have that for some $c>0$ and $x\in[0,1]$,*

$$
\left|\mathbb{E}[Z^x\left(U,\delta\right)] - x\right| \ge c\eta
$$

Proof. We use the notations from the proof of Lemma A.1 such as $Z^x\left(U,\delta\right)$, and denote that $\mathbb{E}\left[Z^x\left(U,\delta\right)\right] \equiv \mathbb{E}_{U,\delta|\boldsymbol{z}}\left[Z^x\left(U,\delta\right)\right]$. Using the proof in Lemma A.2, we have

$$
\begin{aligned}
\mathbb{E}\left[Z^x\left(U,\delta\right)\right] \;=\; & F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) + \int_0^x \Bigg\{ - F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) + F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
& + \left[1-F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right]\frac{x-s}{1-s}\Bigg\} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right). \qquad (9)
\end{aligned}
$$

Furthermore, the right term of Eqn. (9) is

$$
\begin{aligned}
& \int_0^x \Bigg\{ - F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) + F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\left(1-\frac{x-s}{1-s}\right) + \frac{x-s}{1-s}\Bigg\} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
= \;& \int_0^x \Bigg[- \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) - x\right\} + \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\}\frac{1-x}{1-s} \\
& \qquad -x + s\frac{1-x}{1-s} + \frac{x-s}{1-s}\Bigg] dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
= \;& \int_0^x \left[- \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) - x\right\} + \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\}\frac{1-x}{1-s}\right] dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
= \;& -\int_0^x \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x)\right) - x\right\} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) + \int_0^x \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\}\frac{1-x}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right).
\end{aligned}
$$

By the Regularity Condition 2, we can pick x^* for x such that $\left|F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^*\right| = \eta > 0$.

$$
\begin{aligned}
&\left|\mathbb{E}[Z^{x^*}(U,\delta)] - x^*\right| \\
&= \left| F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^* - \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^*\right\} \int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right. \\
&\qquad\qquad \left. + \int_0^{x^*} \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\} \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right| \\
&= \left| \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^*\right\} \left(1 - \int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right) \right. \\
&\qquad\qquad \left. + \int_0^{x^*} \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\} \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right| \\
&\geq \left| \left|F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^*\right| \cdot \left(1 - \int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right) \right. \\
&\qquad\qquad \left. - \left| \int_0^{x^*} \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\} \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right| \right|
\end{aligned}
$$

by the triangle inequality. Note that

$$
\begin{aligned}
&\left| \int_0^{x^*} \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\} \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right| \\
&\leq \int_0^{x^*} \left|F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right| \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
&\leq \int_0^{x^*} \left|F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right| dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \\
&\leq \left|F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(x^*)\right) - x^*\right| = \eta.
\end{aligned}
$$

Therefore, we can let $\left| \int_0^{x^*} \left\{F\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - s\right\} \frac{1-x^*}{1-s} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) \right| = c'\eta$ for some $c' \in (0,1)$. Finally, we have

$$
\left|\mathbb{E}[Z^{x^*}(U,\delta)] - x^*\right| \geq \left|\eta\left(1 - \int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right)\right) - c'\eta\right| = c\eta
$$

where $c = \left|1 - \int_0^{x^*} dG\left(F_{\hat{\boldsymbol{\theta}}}^{-1}(s)\right) - c'\right|$. Furthermore, $c \in (0,1)$ by Regularity Condition 2. This completes the proof. □

Lemma A.4. *Under the Regularity Conditions, we have*

$$
\mathbb{P}\left(\sup_{x\in[0,1]} \left|\tilde{F}(x) - x\right| \geq c\eta/2\right)
$$

goes to 1 as $N \to \infty$.

Proof. By Eqn. (4), replacing x with the $\mathbb{E}[\tilde{F}(x)]$, we have that for any $\delta > 0$,

$$
\mathbb{P}\left(\sup_{s\in[0,1]} \left|\tilde{F}(s) - \mathbb{E}[\tilde{F}(s)]\right| < \delta\right) \to 1
$$

as $N \to \infty$. Lemma A.3 implies $\left|\mathbb{E}[\tilde{F}(x^*)] - x^*\right| \geq c\eta$ for x^* in the Regularity Conditions. Consequently, the probability of $\left|\tilde{F}(x^*) - x^*\right| \geq c\eta/2$ goes to 1 as $N \to \infty$. It suffices for the proof. □

B Likelihood function

B.1 DeepSurv

The classical Cox PH model (Cox, 1972), $\lambda(t \mid \boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{z})$, assumes linearity for the effect of the variables $\boldsymbol{z}$. Here, we can adopt a DNN to estimate the effects of variables beyond linearity. The hazard function is then modeled as $\lambda(t \mid \boldsymbol{z}) = \lambda_0(t)\exp\{g(\boldsymbol{z};\boldsymbol{\theta})\}$ where g is a nonlinear function of $\boldsymbol{z}$ obtained through the DNN. The original Cox PH model optimizes the logarithm of the partial likelihood which is given by

$$\sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta}^\intercal \boldsymbol{z}_i - \log \sum_{j \in R_i} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{z}_j)\right]$$

where $R_i = \{j : Y_j \geq Y_i\}$ is a risk set and $\boldsymbol{\beta}$ is the vector of regression coefficients. Maximizing this partial likelihood or minimizing the negative log-partial likelihood yields the estimated coefficients $\hat{\boldsymbol{\beta}}$. If we adopt DeepSurv (Katzman et al., 2018), the negative log-partial likelihood is changed to $-\sum_{i=1}^{N} \delta_i \left[g(\boldsymbol{z}_i;\boldsymbol{\theta}) - \log \sum_{j \in R_i} \exp\{g(\boldsymbol{z}_j;\boldsymbol{\theta})\}\right]$ for the loss function. In DeepSurv, we only use the estimated $g(\boldsymbol{z}_i;\boldsymbol{\theta})$ to calculate the C-index since the baseline hazard function $\lambda_0(t)$ is equivalent for all subjects if the time t is the same. But to get $\hat{F}_{\boldsymbol{\theta}}$, we need to estimate $\lambda_0(t)$. Rather than estimating itself, we estimate the cumulative baseline hazard function $\Lambda_0(t)$ using the Breslow estimator, denoted by $\hat{\Lambda}_0(t)$ (Lin, 2007). Then, by using the relationship between the distribution function and the cumulative hazard function, we have

$$\hat{F}_{\boldsymbol{\theta}}(t \mid \boldsymbol{z}) = 1 - \exp\left[-\hat{\Lambda}_0(t)\exp\{g(\boldsymbol{z};\hat{\boldsymbol{\theta}})\}\right]$$

It is known that $\hat{\Lambda}_0(t)$ is consistent for $\Lambda_0(t)$. However, $\hat{F}_{\boldsymbol{\theta}}$ includes the effect of $\boldsymbol{z}$, which tells us $\hat{F}_{\boldsymbol{\theta}}$ is not consistent for F, even when $\hat{\Lambda}_0(t)$ is consistent. Therefore, we need to consider the calibration. Instead, the calibration error we observed is lower than that of other models.

B.2 MTLR

MTLR (Yu et al., 2011) models the probability of failure using a logistic function. Let $(t_1, t_2, \ldots, t_B)$ be the previously determined time points, $(y_1, y_2, \ldots, y_B)$ be the survival status according to $(t_1, t_2, \ldots, t_B)$, and $(s_1, s_2, \ldots, s_N)$ be the actual survival times of the subjects. The probability of failure after the l-th time point is modeled as $(1 + \exp(-\boldsymbol{\theta}_l^\intercal \boldsymbol{z}))^{-1}$. When the survival time is observed over the l-th point, the survival time is vectorized into $(\underbrace{0, 0, \ldots, 0}_{\text{\# of } l \text{ points}}, 1, 1, \ldots, 1)$ when $\delta = 1$, and $(\underbrace{0, 0, \ldots, 0}_{\text{\# of } l \text{ points}})$ when $\delta = 0$. For the uncensored case, we obtain the probability of observing the status $(y_1(s_i), y_2(s_i), \ldots, y_B(s_i))$ as

$$f_{\boldsymbol{\theta}}(s_i \mid \boldsymbol{z}_i) = \mathbb{P}_{\boldsymbol{\theta}}(Y = (y_1, y_2, \ldots, y_B) \mid \boldsymbol{z}_i) = \frac{\exp\left\{\sum_{k=1}^{B} y_k(s_i)(\boldsymbol{\theta}_k^\intercal \boldsymbol{z}_i)\right\}}{\sum_{k=0}^{B} \exp\left\{\sum_{j=k+1}^{B}(\boldsymbol{\theta}_k^\intercal \boldsymbol{z}_i)\right\}}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B)^\intercal$.

For the censored case, we only know that the failure would be observed after the censoring time s_i that we have observed. Thus, if we denote t_l as the closest time point following s_i, the likelihood of a

censored subject is given by

$$S_{\boldsymbol{\theta}}\left(s_i \mid \boldsymbol{z}_i\right)=\mathbb{P}_{\boldsymbol{\theta}}\left(T_i \geq t_l \mid \boldsymbol{z_i}\right)=\frac{\sum_{k=l}^{B} \exp\left\{\sum_{j=k+1}^{B}\left(\boldsymbol{\theta}_j^{\intercal} \boldsymbol{z}_i\right)\right\}}{\sum_{k=0}^{B} \exp\left\{\sum_{j=k+1}^{B}\left(\boldsymbol{\theta}_j^{\intercal} \boldsymbol{z}_i\right)\right\}}.$$

By combining these, the negative log-likelihood for the loss function is defined as

$$\ell(\boldsymbol{\theta})=-\sum_{i=1}^{N}\left[\delta_i \log f_{\boldsymbol{\theta}}\left(s_i \mid \boldsymbol{z}_i\right)+\left(1-\delta_i\right) \log S_{\boldsymbol{\theta}}\left(s_i \mid \boldsymbol{z}_i\right)\right].$$

Here, $f_{\boldsymbol{\theta}}, S_{\boldsymbol{\theta}}$ are the conventional notations, not the true probability density function and survival function of failure time. For more details, refer to Yu et al. (2011). For calibration error, $\hat{F}_{\boldsymbol{\theta}}$ would be obtained, which is accomplished by

$$\hat{F}_{\boldsymbol{\theta}}\left(t \mid \boldsymbol{z}\right)=\frac{\sum_{k=0}^{l} \exp\left\{\sum_{j=k+1}^{B} \hat{\boldsymbol{\theta}}_j^{\intercal} \boldsymbol{z}\right\}}{\sum_{k=0}^{B} \exp\left\{\sum_{j=k+1}^{B} \hat{\boldsymbol{\theta}}_j^{\intercal} \boldsymbol{z}\right\}} \tag{10}$$

where $t \in [t_l, t_{l+1})$. We use $B = 20$ for all experiments, except for SEER-stomach and SEER-lung, where $B = 10$ is used.

B.3 Parametric model

We use the parametric model proposed by Goldstein et al. (2020), referred to as LognormalNN. We parametrize the location and scale parameters (μ, σ) of the lognormal distribution. Let f and S be the probability density function and survival function of the lognormal distribution, respectively, with parameters estimated by DNN. The negative log-likelihood is as $\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\left[\delta_i \log f_{\boldsymbol{\theta}}\left(y_i \mid \boldsymbol{z}_i\right)+\left(1-\delta_i\right) \log S_{\boldsymbol{\theta}}\left(y_i \mid \boldsymbol{z}_i\right)\right]$ where $\boldsymbol{\theta} = (\mu, \sigma)^{\intercal}$ and y_i is an observed time of subject i. Then, we calculate the distribution function with two parameters at the given observed time.

B.4 CRPS

Continuous Ranked Probability Score (CRPS; Avati et al., 2020) aims to minimize

$$\int_0^y\left\{\hat{F}_{\boldsymbol{\theta}}\left(t \mid \boldsymbol{z}\right)\right\}^2 dt+\delta \int_y^{\infty}\left\{1-\hat{F}_{\boldsymbol{\theta}}\left(t \mid \boldsymbol{z}\right)\right\}^2 dt$$

where y is an observed time and δ is a censoring indicator. See Appendix B in Avati et al. (2020) for a lognormal distribution. Since there is a closed form for the loss, we use the lognormal distribution for $\hat{F}_{\boldsymbol{\theta}}$.

B.5 DeepHit

DeepHit (Lee et al., 2018) is a discrete time survival model that estimates the probability mass function $\mathbb{P}\left(T = t \mid \boldsymbol{z}\right)$ over discrete time intervals. Let $T_{\max}$ denote the maximum time point. Given $\boldsymbol{z}$, the model outputs a probability vector $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{T_{\max}})$ with $\sum_{t=1}^{T_{\max}} \hat{p}_t = 1$ using the softmax function. The model is trained to minimize the negative log-likelihood. For an uncensored sample with event time t, the loss is $-\log \hat{p}_t$, and for a censored sample with censoring time t, the loss is $-\log \sum_{s=t+1}^{T_{\max}} \hat{p}_s$, corresponding to the estimated survival probability $\hat{S}\left(t \mid \boldsymbol{z}\right)$. We adopt DeepHit in the single-event setting and train it using only the log-likelihood loss, without the ranking loss originally proposed in Lee et al. (2018). Similar to MTLR, DeepHit depends on the number of bins, and we use a different number of bins depending on the dataset.

B.6 Weibull AFT model

The Weibull AFT model (Stute, 1993) assumes the following regression form:

$$\log T = \boldsymbol{\beta}^{\intercal} \boldsymbol{z} + \sigma \epsilon$$

where ϵ follows a standard Gumbel distribution. Under this formulation, T follows a Weibull distribution with a scale parameter $\exp(\boldsymbol{\beta}^\intercal \boldsymbol{z})$ and a shape parameter $1/\sigma$. The survival function is given by:

$$S(t \mid \boldsymbol{z}) = \exp\left(-\left\{t \exp\left(-\boldsymbol{\beta}^\intercal \boldsymbol{z}\right)\right\}^{1/\sigma}\right).$$

The likelihood function is defined similarly to the parametric model.

C Proof for Proposition 5.1

In this section, we demonstrate that if survival curves do not cross, the KSP preserves the ordering of mean survival times. Furthermore, we show that this property holds for both DeepSurv and the Weibull AFT model after applying KSP.

First, recall that for two individuals with variables $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, if

$$F(s \mid \boldsymbol{z}_i) < F(s \mid \boldsymbol{z}_j) \quad \text{for } s \geq 0,$$

then the CDFs after the KSP also satisfy $F^*(s \mid \boldsymbol{z}_i) < F^*(s \mid \boldsymbol{z}_j)$ due to the monotonicity of the transformation.

For a non-negative random variable T, it is known that its expectation can be written as

$$\mathbb{E}[T] = \int_0^\infty \{1 - F(s)\}\, ds.$$

Applying this identity to both the original and the transformed CDFs, we obtain

$$\begin{aligned} & \mathbb{E}[T \mid \boldsymbol{z}_i] = \int_0^\infty [1 - F(s \mid \boldsymbol{z}_i)]\, ds > \int_0^\infty [1 - F(s \mid \boldsymbol{z}_j)]\, ds = \mathbb{E}[T \mid \boldsymbol{z}_j] \\ \iff \quad & \mathbb{E}[T^* \mid \boldsymbol{z}_i] > \mathbb{E}[T^* \mid \boldsymbol{z}_j]. \end{aligned}$$

where $T^* \sim F^*$. Thus, the ordering of the mean survival times is preserved; consequently, the C-index based on the mean survival time remains unchanged after KSP.

DeepSurv Under the PH assumption, the ordering of mean survival times is preserved. The model assumes that

$$\begin{aligned} & T_i < T_j \\ \iff & \exp\{g(\boldsymbol{z}_i; \boldsymbol{\theta})\} > \exp\{g(\boldsymbol{z}_j; \boldsymbol{\theta})\} \\ \iff & \exp\left[-\int_0^t \lambda_0(s) \exp\{g(\boldsymbol{z}_i; \boldsymbol{\theta})\} ds\right] < \exp\left[-\int_0^t \lambda_0(s) \exp\{g(\boldsymbol{z}_j; \boldsymbol{\theta})\} ds\right],\ \forall t \geq 0 \\ \iff & \mathbb{E}[T \mid \boldsymbol{z}_i] < \mathbb{E}[T \mid \boldsymbol{z}_j] \\ \iff & \mathbb{E}[T^* \mid \boldsymbol{z}_i] < \mathbb{E}[T^* \mid \boldsymbol{z}_j] \end{aligned}$$

where $g(\boldsymbol{z}; \boldsymbol{\theta})$ is the log-risk function. Hence, the C-index derived from the risk scores is preserved under KSP.

Weibull AFT The Weibull AFT model defines the survival function as

$$S(t \mid \boldsymbol{z}) = \exp\left(-\left\{t \exp\left(-\boldsymbol{\beta}^\intercal \boldsymbol{z}\right)\right\}^{1/\sigma}\right).$$

And,

$$\begin{aligned} & T_i < T_j \\ \iff & \exp(\boldsymbol{\beta}^\intercal \boldsymbol{z}_i) < \exp(\boldsymbol{\beta}^\intercal \boldsymbol{z}_j) \\ \iff & S(t \mid \boldsymbol{z}_i) < S(t \mid \boldsymbol{z}_j),\ \forall\, t \geq 0 \\ \iff & \mathbb{E}[T \mid \boldsymbol{z}_i] < \mathbb{E}[T \mid \boldsymbol{z}_j] \\ \iff & \mathbb{E}[T^* \mid \boldsymbol{z}_i] < \mathbb{E}[T^* \mid \boldsymbol{z}_j] \end{aligned}$$

Therefore, the expected survival times maintain their original ordering.

Table 3: Summary of the ten real-world datasets.

Group	Dataset	# Samples (Train / Validation / Test)	% Censored	# Covariates
Small	WHAS	982 / 328 / 328	58%	6
	METABRIC	1,142 / 381 / 381	42%	9
	GBSG	1,340 / 446 / 446	43%	7
	NACD	1,436 / 480 / 480	36%	51
Medium	NB-SEQ	2,873 / 958 / 958	31%	24
	SUPPORT	5,325 / 1,774 / 1,774	32%	14
	MIMIC-III	5,353 / 1,785 / 1,785	0%	15
Large	SEER-liver	16,575 / 5,525 / 5,526	22%	15
	SEER-stomach	20,615 / 6,872 / 6,872	27%	15
	SEER-lung	125,254 / 41,751 / 41,752	16%	15

D Data generation and description

We first summarize the ten datasets used in this study. Then, we describe each dataset in detail, along with the corresponding pre-processing steps.

D.1 Details for WHAS, METABRIC, GBSG, and SUPPORT

The DeepSurv package (Katzman et al., 2018) in Python includes four pre-processed real-world datasets: the Worcester Heart Attack Study (WHAS), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), German Breast Cancer Study Group (GBSG), and the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT). To ensure consistency across experiments, we combined the training and test sets provided by the package and re-split the data into training, validation, and test sets in a 3:1:1 ratio. No additional data pre-processing was performed.

D.2 Details for NACD

The Northern Alberta Cancer Dataset (NACD) (Qi et al., 2024a) includes patients diagnosed with various cancers, such as lung, colorectal, head and neck, esophageal, stomach, and other types of cancer. The event of interest in this dataset is failure time. We obtain the dataset from http://pssp.srv.ualberta.ca under the “Public Predictors" section. It consists of 2,396 patients with 51 features. For our experiments, we split the dataset into training, validation, and test sets (1,436 / 480 / 480), ensuring that the censoring rate remains balanced at 36%.

D.3 Details for NB-SEQ

NB-SEQ consists of neuroblastoma data and other supervised penalty learning benchmarks specifically designed for censored regression in the context of supervised penalty function learning for change-point detection. The dataset includes various summary statistics, as well as the minimum and maximum values of λ, achieving high performance, where λ is a hyper-parameter for the penalty function. The primary response variable (considered a failure time) is the maximum value of the estimated λ. It is important to note that the maximum value of λ is not directly observed in some cases; we only have information that the value exceeds a certain threshold, indicating that the true value is censored. We analyze 4,789 sequence data points with 24 variables. The list of the 24 variables used in the NB-SEQ dataset is provided in Table 4. All the variables are continuous variables. The censoring rate is 31%. The dataset is partitioned into sizes 2,873, 958, and 958 subsets.

D.4 Details for MIMIC-III

MIMIC-III consists of de-identified health-related information from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA, with a primary focus on intensive care unit (ICU) admissions. Following the pre-processing steps introduced in Harutyunyan et al. (2019), provided by https://github.com/YerevaNN/mimic3-benchmarks. This dataset involves

Table 4: Twenty four variables in NB-SEQ

quartile.25. / quartile.50 / quartile.75 / quartile.100 / mean / sd / bases / data / log.1.quartile.25 / log.1.quartile.50 / log.1.quartile.75 / log.1.quartile.100 / log.1.mean / log.1.sd / log.1.bases / log.1.data / log.quartile.100 / log. mean / log.sd / log.bases / log.data / log.log.quartile.100 / log.log.bases / log.log.data

multiple instances of length-of-stay data, with each subject having repeated measures at different time points. After excluding ICU transfers and patients under 18, the training and test sets consist of 2,925,434 and 525,912 instances, respectively. Recognizing that using the provided data directly is not suitable due to the correlation among data points for a single subject, we undertake additional steps. First, we exclude variables with excessively high missing rates. Next, we create one instance per subject by defining survival time as the duration between enrollment and discharge, which has no censoring. The dataset includes a list of 15 variables, as shown in Table 5. For variables that are repeatedly measured, we transform them into a single representative value, such as the mean for continuous variables and the mode for categorical variables. Finally, we exclude subjects whose variables are still missing, even after applying imputation, resulting in a dataset with 8,873 instances that combines the training set with the test set. We partition the entire dataset into training, validation, and test sets, with sizes of 5,353, 1,785, and 1,785, respectively.

Table 5: Fifteen selected variables in MIMIC-III.

Name	Type
Diastolic blood pressure	Continuous
Glascow coma scale eye opening	Categorical
Glascow coma scale motor response	Categorical
Glascow coma scale total	Continuous
Glascow coma scale verbal response	Categorical
Glucose	Continuous
Heart Rate	Continuous
Height	Continuous
Mean blood pressure	Continuous
Oxygen saturation	Continuous
Respiratory rate	Continuous
Systolic blood pressure	Continuous
Temperature	Continuous
Weight	Continuous
pH	Continuous

D.5 Details for SEER-liver, SEER-stomach, and SEER-lung

The Surveillance, Epidemiology, and End Results (SEER) Program dataset (National Cancer Institute, DCCPS, Surveillance Research Program, 2015) is a large-scale, population-based cancer registry that covers approximately 49% of the U.S. population (Qi et al., 2024a). It contains comprehensive information on cancer incidence, patient demographics, treatments, and survival outcomes. In this study, we utilize three distinct subsets from the SEER database—SEER-liver, SEER-stomach, and SEER-lung—which respectively include patients diagnosed with liver, stomach, and lung cancers. The objective is to model the time from cancer diagnosis to a failure event, specifically death, with time measured in months. For our analysis, we selected a subset of clinically relevant variables: Sex, Race, Summary stage, Malignant behavior, Record number, Total tumor size, and Age at diagnosis. Variables with high missing rates or a large number of categories were excluded, and patients without recorded follow-up (due to immediate death) were removed. Additionally, we applied filtering criteria to exclude patients with extreme or implausible values. Specifically, we retained only patients with a record number less than or equal to 8, a total tumor size less than or equal to 13, and a survival time greater than or equal to 10 months. The dataset can be downloaded from https://seer.cancer.gov/.

E Comparison of post-processing methods: Figure

In this section, we present the results across real datasets using figures. The error bars represent 95% confidence intervals.

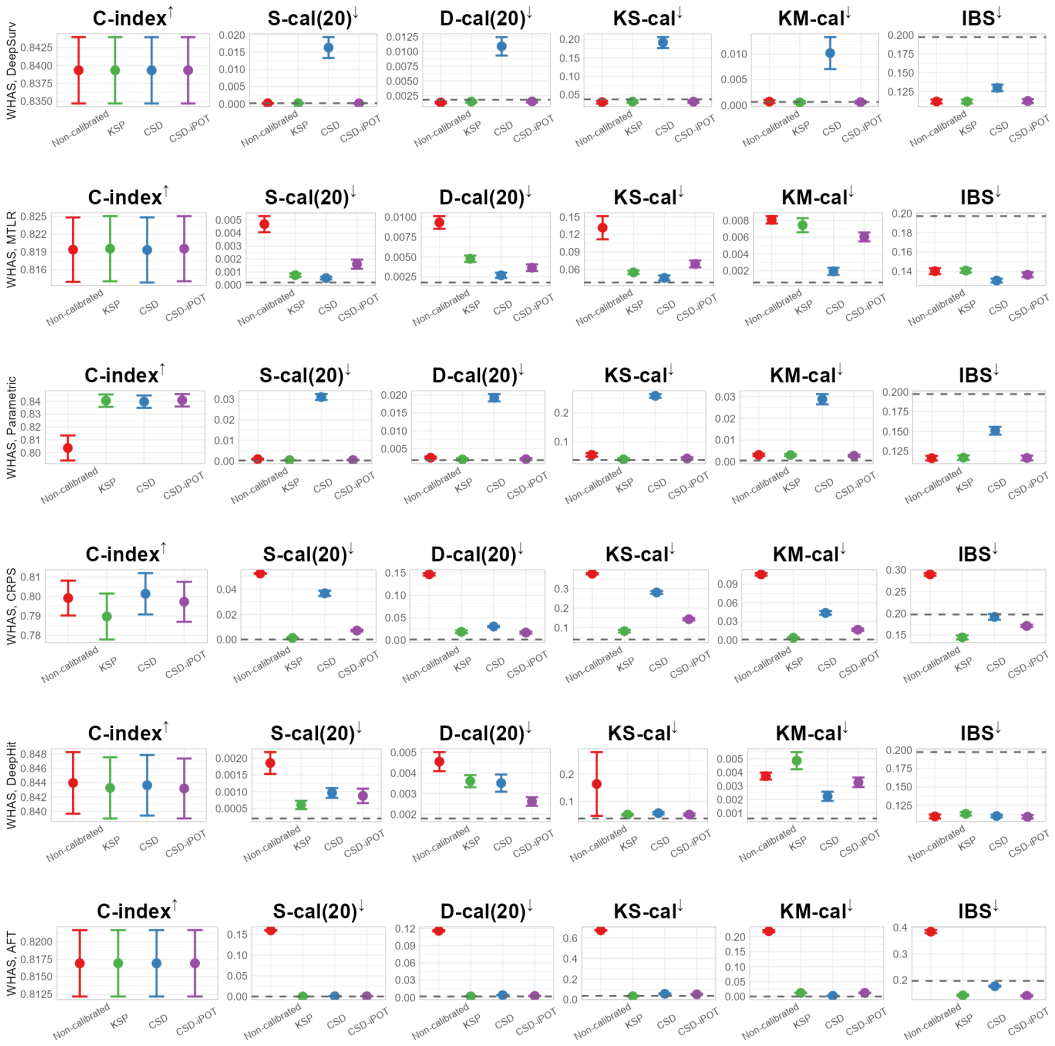


Figure 5: Comparison of post-processing methods for the WHAS dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

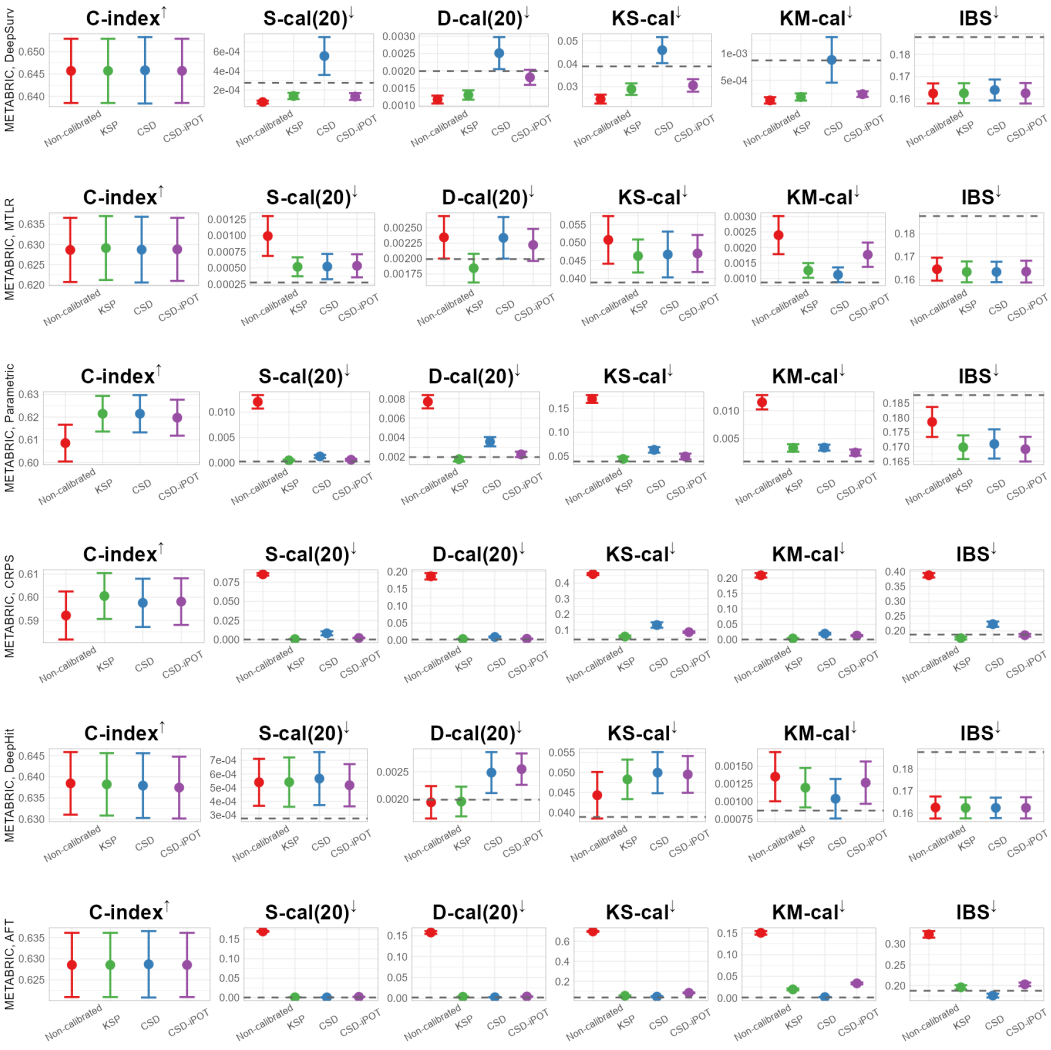


Figure 6: Comparison of post-processing methods for the METABRIC dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

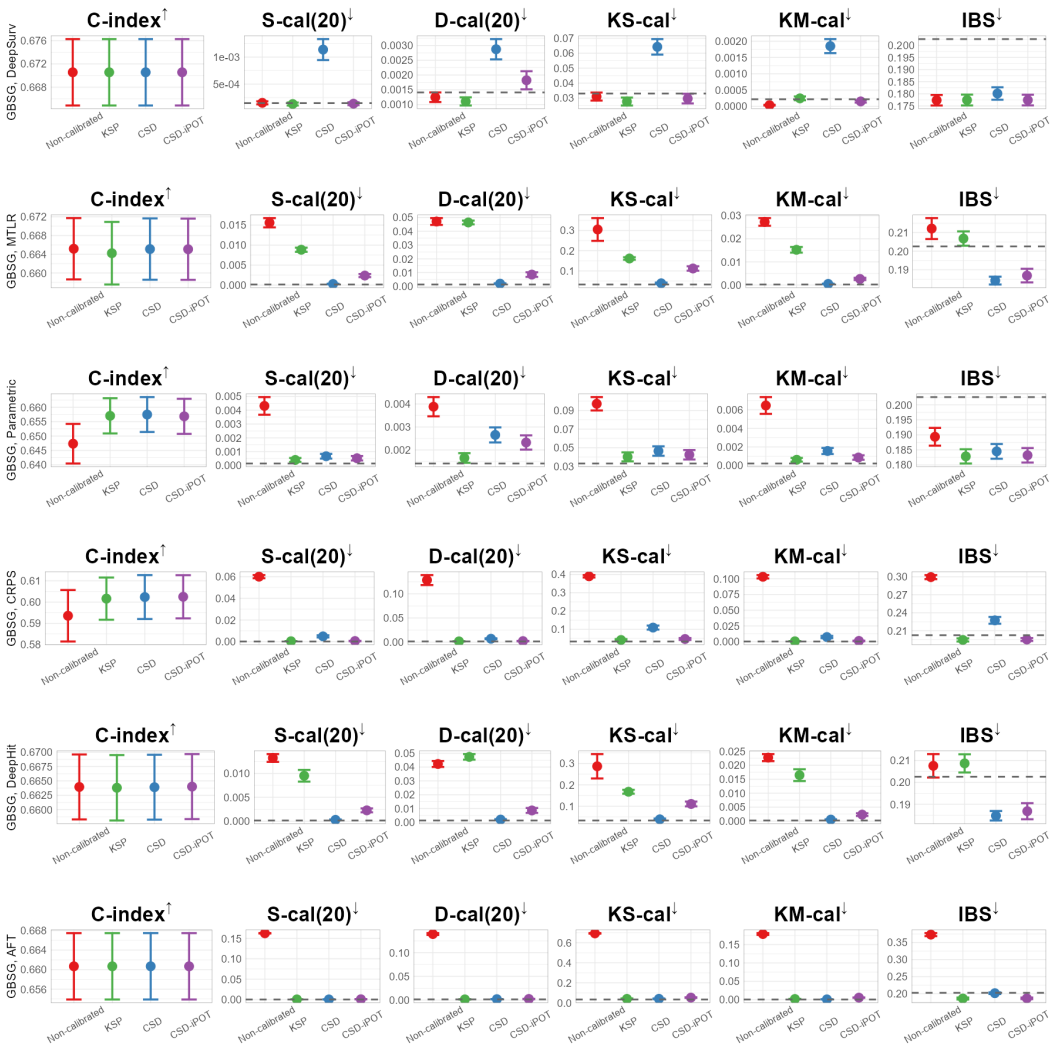


Figure 7: Comparison of post-processing methods for the GBSG dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

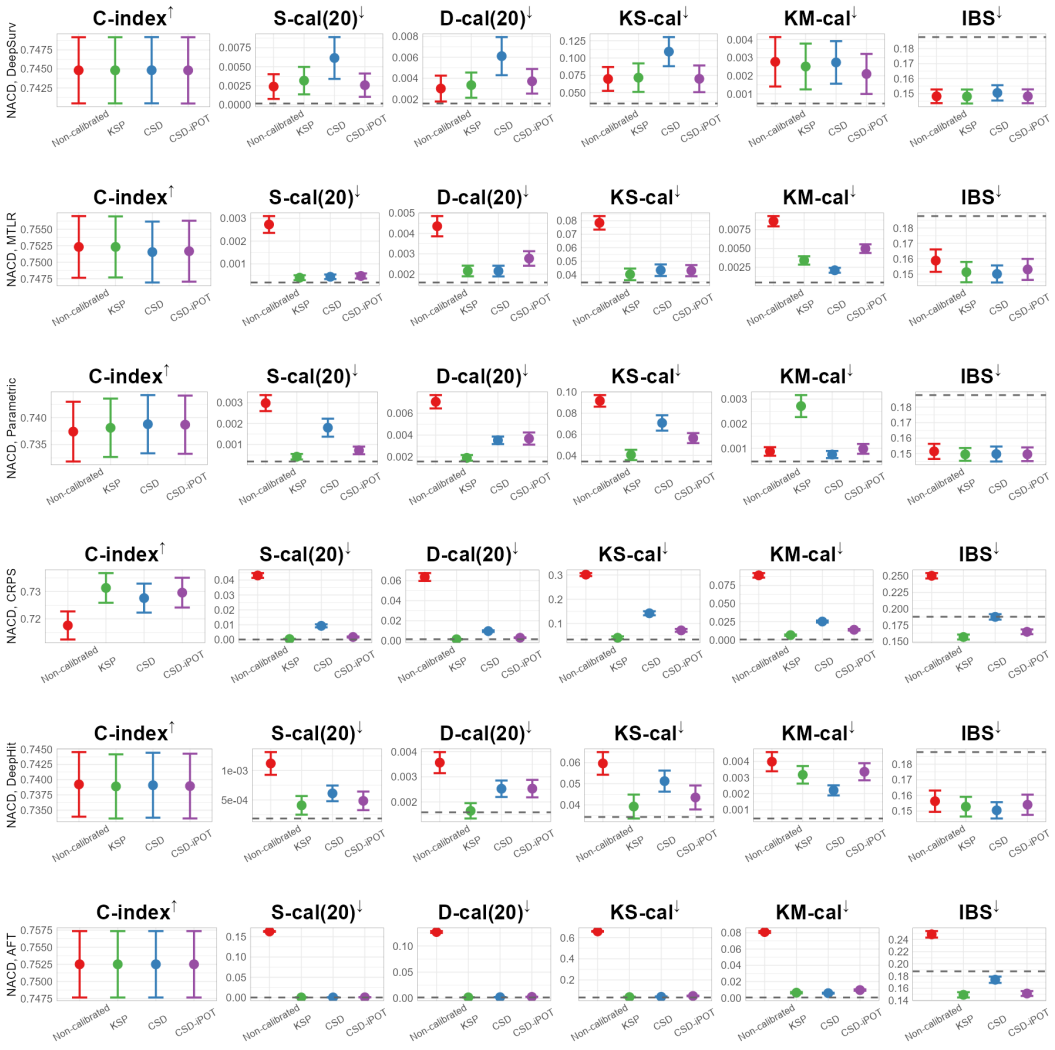


Figure 8: Comparison of post-processing methods for the NACD dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

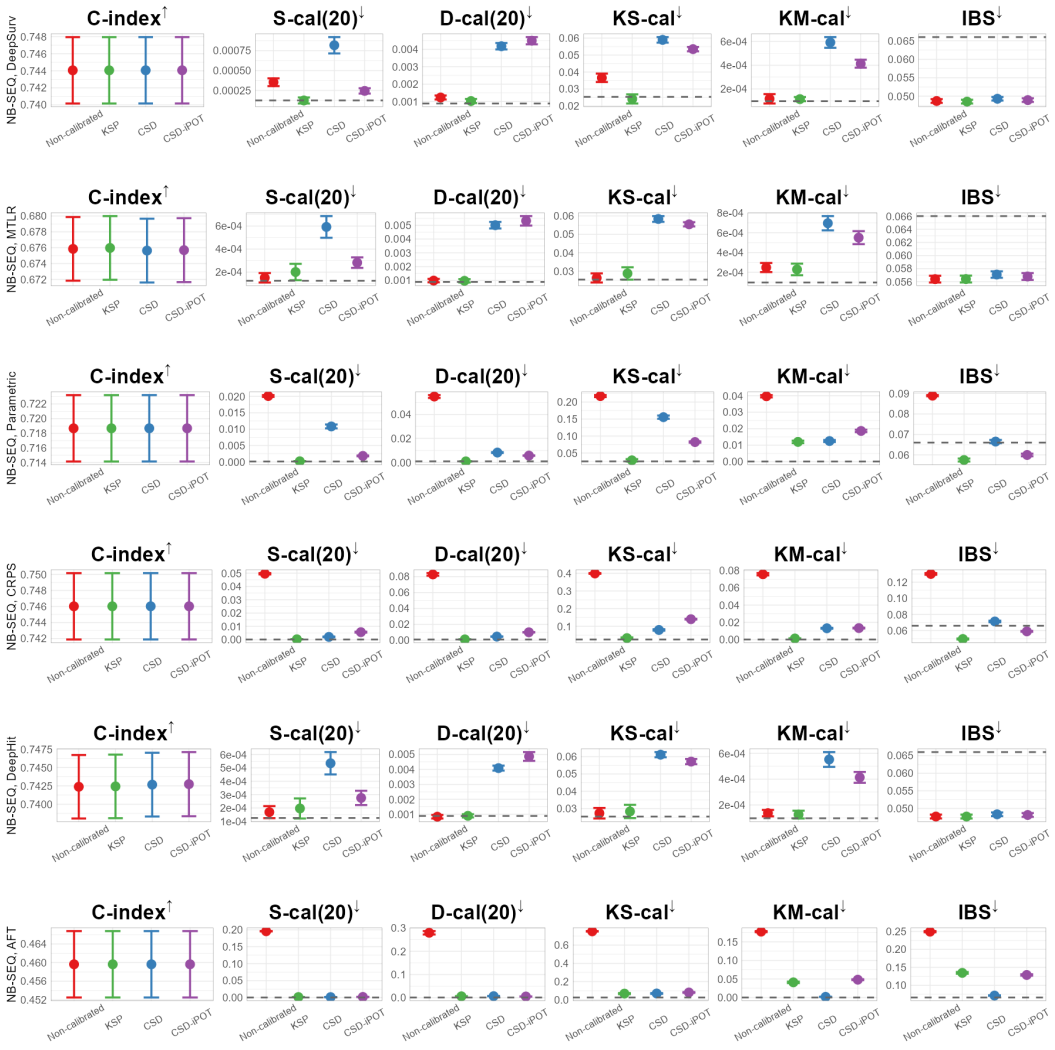


Figure 9: Comparison of post-processing methods for the NB-SEQ dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

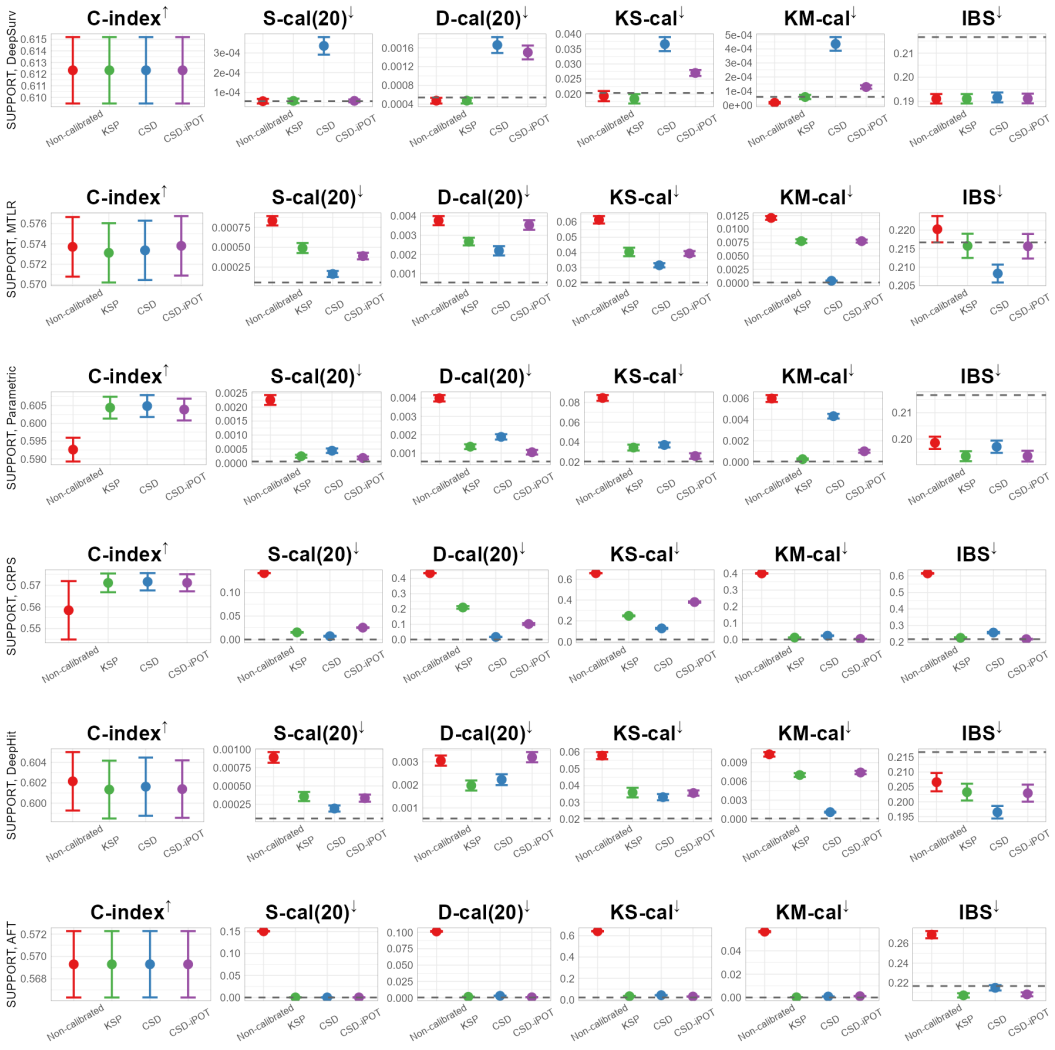


Figure 10: Comparison of post-processing methods for the SUPPORT dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

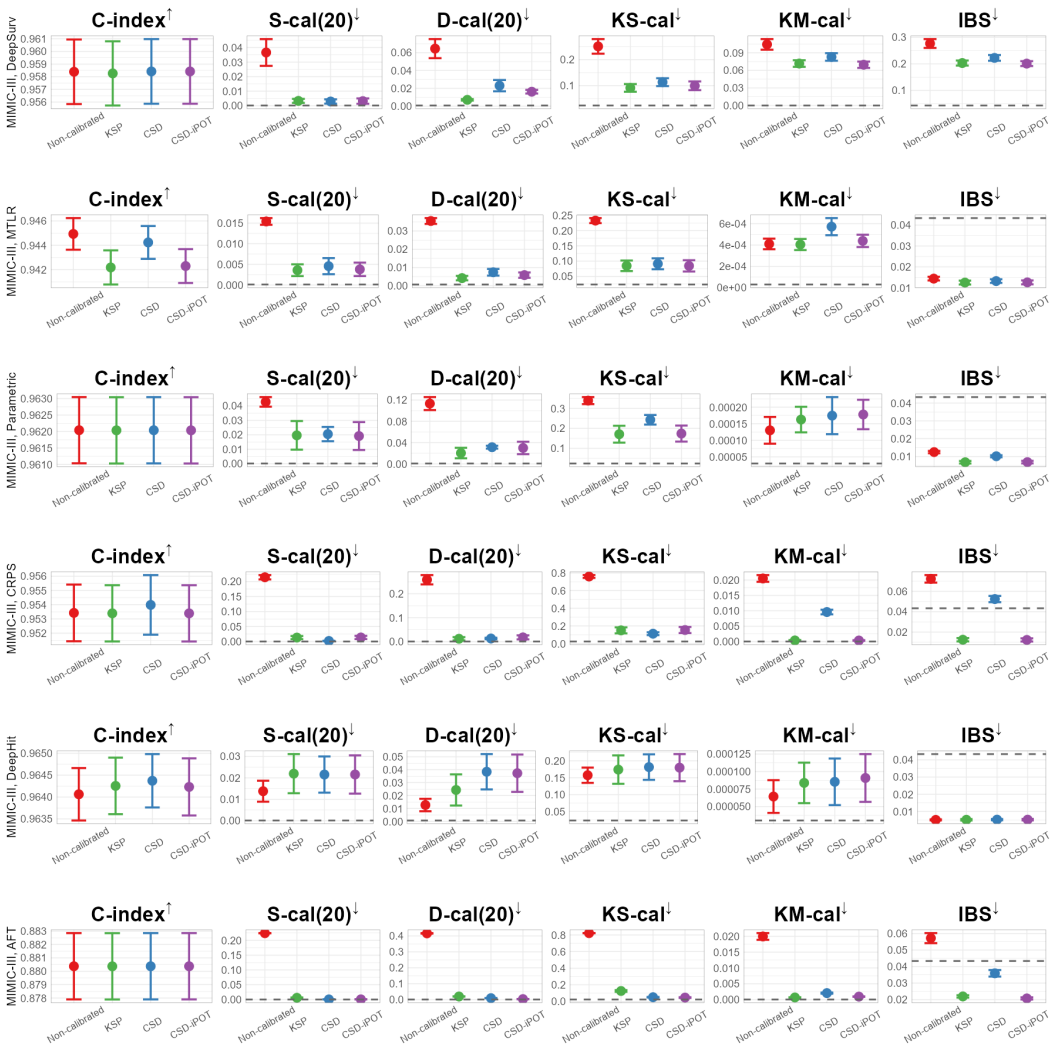


Figure 11: Comparison of post-processing methods for the MIMIC-III dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

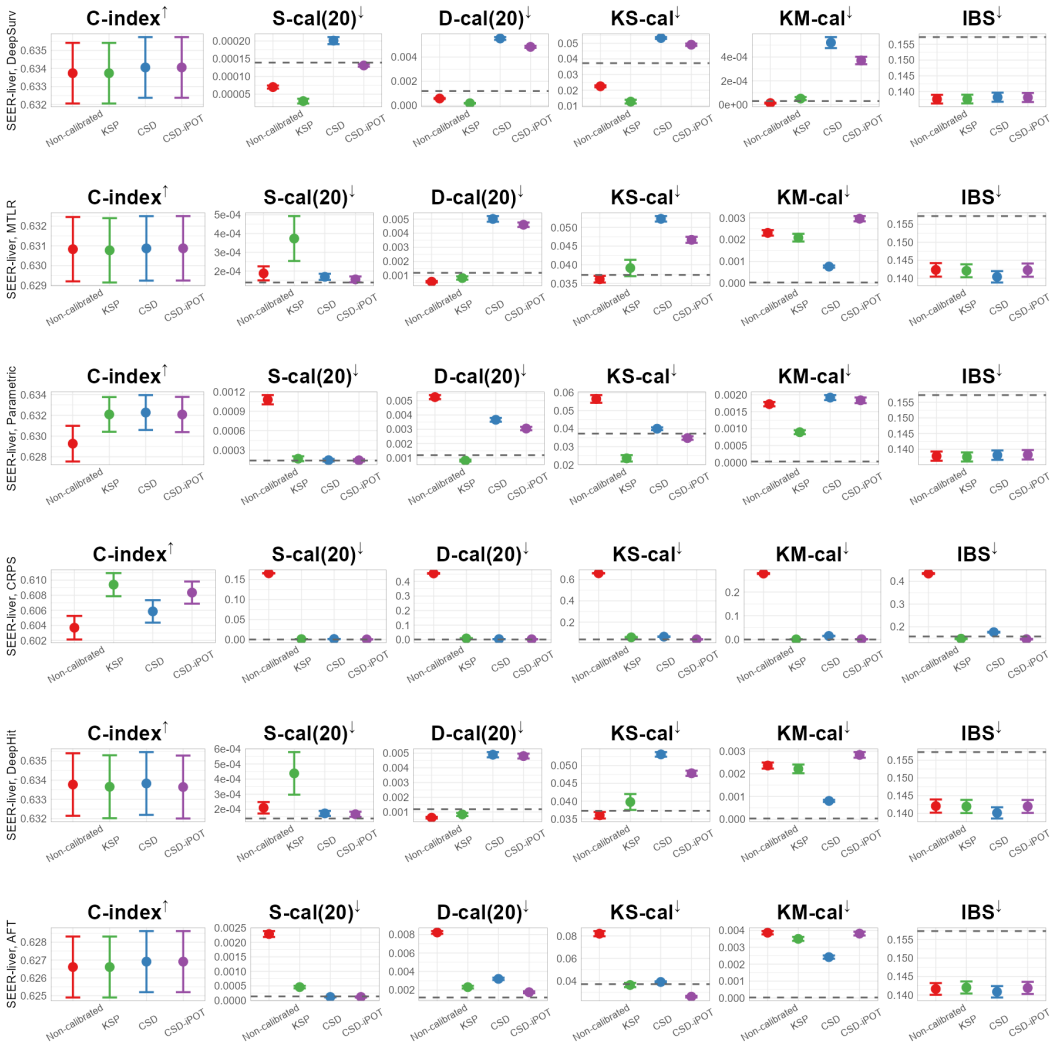


Figure 12: Comparison of post-processing methods for the SEER-liver dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

Figure 13: Comparison of post-processing methods for the SEER-stomach dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

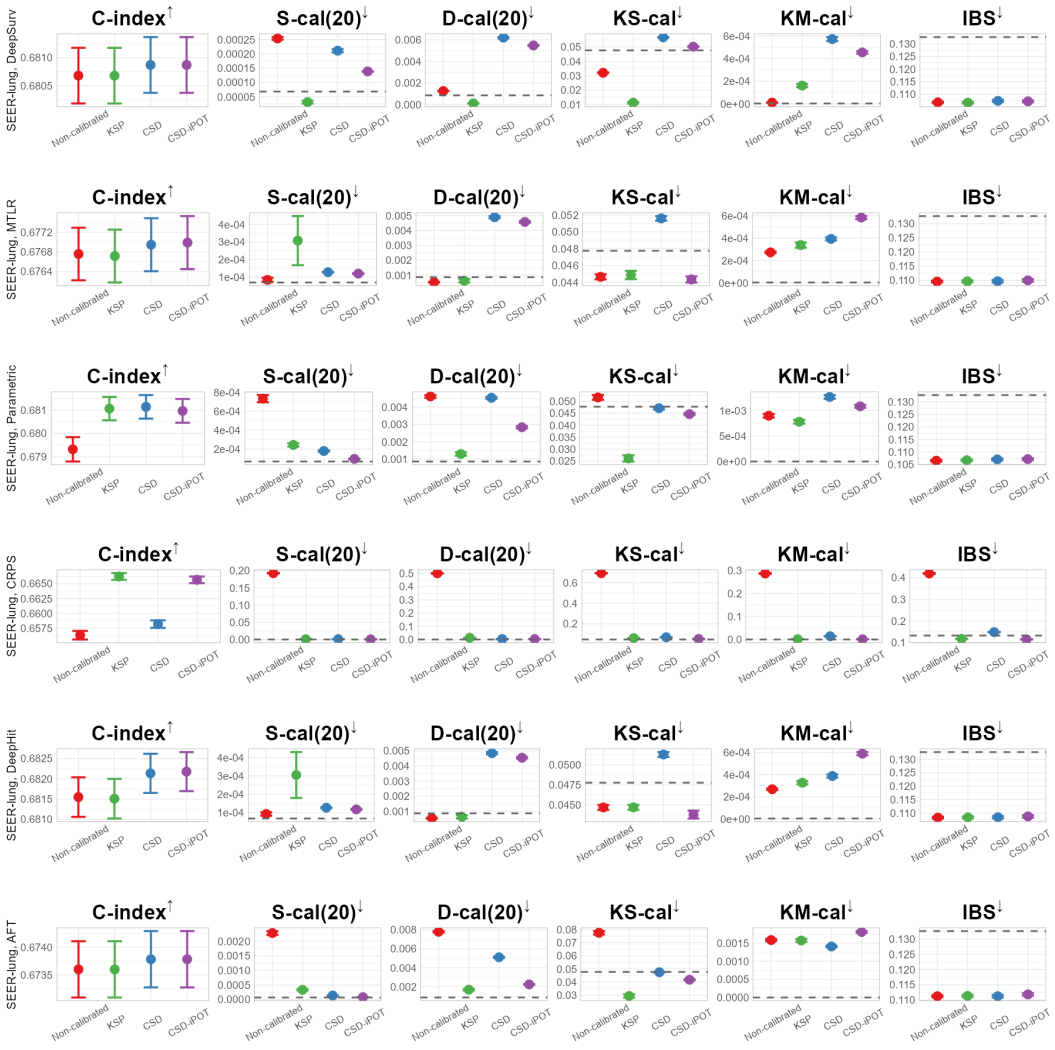


Figure 14: Comparison of post-processing methods for the SEER-lung dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all calibration metrics. The dashed line indicates the calibration error of the KM estimator.

F Comparison of post-processing methods: Table

In this section, we present the results in tabular form, in contrast to Appendix E, where the same results are visualized as figures. Similar to Table 1 in the main text, we summarize the number of cases in which each post-processing method outperforms the others for each model. We implemented X-cal (Goldstein et al., 2020 in the main paper), CSD (Qi et al., 2024a), and CSD-iPOT (Qi et al., 2024b) with reference to publicly available code repositories under the MIT License: `https://github.com/rajesh-lab/X-CAL`, `https://github.com/shi-ang/CSD`, and `https://github.com/shi-ang/MakeSurvivalCalibratedAgain`.

Table 6: Summary of pairwise comparisons between post-processing methods for DeepSurv. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepSurv	KSP	0 (0)	**6 (5)**	**6 (5)**	**7 (6)**	4 (1)	**6 (1)**
	Non-calibrated	1 (0)	3 (1)	4 (0)	3 (1)	**6 (5)**	4 (0)
	Ties	**9**	1	0	0	0	0
	KSP	0 (0)	**9 (9)**	**10 (10)**	**10 (10)**	**10 (9)**	**10 (5)**
	CSD	**6 (0)**	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Ties	4	0	0	0	0	0
	KSP	1 (0)	**6 (4)**	**10 (8)**	**8 (5)**	**7 (5)**	**7 (1)**
	CSD-iPOT	**5 (0)**	4 (0)	0 (0)	2 (0)	3 (1)	3 (0)
	Ties	4	0	0	0	0	0

Table 7: Summary of pairwise comparisons between post-processing methods for MTLR. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
MTLR	KSP	3 (0)	**6 (6)**	**7 (5)**	**6 (5)**	**9 (6)**	**7 (2)**
	Non-calibrated	**6 (1)**	4 (3)	3 (2)	4 (2)	1 (1)	3 (0)
	Ties	1	0	0	0	0	0
	KSP	4 (0)	4 (1)	**6 (6)**	**7 (4)**	3 (3)	3 (1)
	CSD	**6 (1)**	**6 (6)**	4 (3)	3 (3)	**7 (6)**	**7 (4)**
	Ties	0	0	0	0	0	0
	KSP	4 (0)	5 (2)	**8 (8)**	**6 (4)**	**7 (6)**	**6 (0)**
	CSD-iPOT	**6 (0)**	**5 (5)**	2 (2)	4 (2)	3 (2)	4 (2)
	Ties	0	0	0	0	0	0

Table 8: Summary of pairwise comparisons between post-processing methods for Parametric model. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
Parametric	KSP	**8 (7)**	**10 (10)**	**10 (10)**	**10 (10)**	**8 (6)**	**8 (5)**
	Non-calibrated	1 (0)	0 (0)	0 (0)	0 (0)	2 (1)	2 (0)
	Ties	1	0	0	0	0	0
	KSP	3 (0)	**8 (7)**	**10 (10)**	**10 (9)**	**9 (6)**	**10 (4)**
	CSD	**5 (0)**	2 (1)	0 (0)	0 (0)	1 (1)	0 (0)
	Ties	2	0	0	0	0	0
	KSP	**6 (0)**	5 (2)	**9 (7)**	**9 (5)**	**7 (5)**	**8 (2)**
	CSD-iPOT	2 (0)	**5 (3)**	1 (1)	1 (1)	3 (2)	2 (0)
	Ties	2	0	0	0	0	0

Table 9: Summary of pairwise comparisons between post-processing methods for CRPS. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
CRPS	KSP	**7 (5)**	**10 (10)**	**10 (10)**	**10 (10)**	**10 (10)**	**10 (10)**
	Non-calibrated	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Ties	1	0	0	0	0	0
	KSP	**5 (2)**	**8 (8)**	**6 (5)**	**8 (8)**	**10 (10)**	**10 (10)**
	CSD	4 (0)	2 (2)	4 (4)	2 (2)	0 (0)	0 (0)
	Ties	1	0	0	0	0	0
	KSP	**5 (0)**	**7 (6)**	5 (3)	**7 (5)**	**6 (6)**	**5 (4)**
	CSD-iPOT	3 (0)	3 (3)	**5 (4)**	3 (3)	4 (3)	**5 (4)**
	Ties	2	0	0	0	0	0

Table 10: Summary of pairwise comparisons between post-processing methods for DeepHit. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepHit	KSP	2 (0)	4 (4)	3 (3)	4 (4)	**7 (4)**	**6 (0)**
	Non-calibrated	**8 (0)**	**6 (3)**	**7 (4)**	**6 (2)**	3 (2)	4 (1)
	Ties	0	0	0	0	0	0
	KSP	1 (0)	4 (3)	**8 (6)**	**8 (6)**	3 (2)	4 (0)
	CSD	**9 (1)**	**6 (5)**	2 (1)	2 (1)	**7 (6)**	**6 (4)**
	Ties	0	0	0	0	0	0
	KSP	5 (0)	3 (2)	**8 (7)**	**7 (3)**	**8 (5)**	**6 (0)**
	CSD-iPOT	**5 (1)**	**7 (4)**	2 (2)	3 (2)	2 (2)	4 (2)
	Ties	0	0	0	0	0	0

Table 11: Summary of pairwise comparisons between post-processing methods for Weibull AFT model. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
AFT	KSP	0 (0)	**10 (10)**	**10 (10)**	**10 (10)**	**9 (8)**	**7 (7)**
	Non-calibrated	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	3 (0)
	Ties	**10**	0	0	0	0	0
	KSP	0 (0)	3 (1)	**8 (8)**	**8 (5)**	2 (2)	**5 (5)**
	CSD	4 (0)	**7 (5)**	2 (2)	2 (2)	**8 (6)**	5 (2)
	Ties	**6**	0	0	0	0	0
	KSP	0 (0)	**6 (5)**	**6 (6)**	**7 (7)**	**10 (9)**	**6 (2)**
	CSD-iPOT	4 (0)	4 (4)	4 (4)	3 (3)	0 (0)	4 (2)
	Ties	**6**	0	0	0	0	0

Table 12: Comparison of post-processing methods for the WHAS dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.83934	0.000184	**0.001324**	**0.029978**	0.000664	**0.111616**
	KSP	0.83934	0.000189	0.001483	0.031463	**0.000526**	0.111655
	CSD	0.83934	0.016320	0.010881	0.191019	0.010207	0.129705
	CSD-iPOT	0.83934	**0.000163**	0.001513	0.031388	0.000563	0.112224
MTLR	Non-calibrated	0.81940	0.004684	0.009308	0.131427	0.008072	0.140125
	KSP	**0.81956**	0.000766	0.004767	0.055003	0.007428	0.140783
	CSD	0.81935	**0.000540**	**0.002696**	**0.045507**	**0.001938**	**0.130369**
	CSD-iPOT	0.81955	0.001610	0.003647	0.069353	0.006033	0.136362
Parametric	Non-calibrated	0.80364	0.000858	0.002488	0.055345	0.003171	**0.116109**
	KSP	0.84049	**0.000379**	**0.001994**	**0.039706**	0.003100	0.116749
	CSD	0.83970	0.031220	0.019276	0.258418	0.028841	0.150842
	CSD-iPOT	**0.84078**	0.000439	0.002074	0.042612	**0.002768**	0.116346
CRPS	Non-calibrated	0.79922	0.052393	0.146602	0.377683	0.105859	0.289911
	KSP	0.78974	**0.001444**	0.018794	**0.081137**	**0.003376**	**0.144378**
	CSD	**0.80142**	0.036708	0.030670	0.280881	0.043592	0.191666
	CSD-iPOT	0.79730	0.007258	**0.017252**	0.142694	0.016375	0.170867
DeepHit	Non-calibrated	**0.84397**	0.001856	0.004547	0.163803	0.003720	0.109808
	KSP	0.84328	**0.000604**	0.003602	**0.051873**	0.004851	0.113300
	CSD	0.84364	0.000963	0.003509	0.057144	**0.002235**	0.110331
	CSD-iPOT	0.84319	0.000875	**0.002621**	0.052076	0.003261	**0.109295**
AFT	Non-calibrated	0.81692	0.159936	0.115956	0.672936	0.218693	0.384233
	KSP	0.81692	**0.000211**	**0.001899**	**0.034510**	0.012560	0.142276
	CSD	**0.81693**	0.000800	0.003983	0.057077	**0.002907**	0.177157
	CSD-iPOT	**0.81693**	0.000744	0.002625	0.051983	0.012709	**0.140417**

Table 13: Comparison of post-processing methods for the METABRIC dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.64572	**0.000074**	**0.001168**	**0.024570**	**0.000129**	**0.162487**
	KSP	0.64572	0.000138	0.001300	0.028909	0.000192	0.162597
	CSD	**0.64585**	0.000554	0.002511	0.045971	0.000882	0.164027
	CSD-iPOT	0.64573	0.000132	0.001810	0.030531	0.000244	0.162564
MTLR	Non-calibrated	0.62866	0.000991	0.002344	0.050756	0.002398	0.164533
	KSP	**0.62912**	**0.000517**	**0.001843**	**0.046266**	0.001261	0.163358
	CSD	0.62873	0.000520	0.002336	0.046683	**0.001120**	**0.163325**
	CSD-iPOT	0.62880	0.000532	0.002220	0.046963	0.001767	0.163472
Parametric	Non-calibrated	0.60856	0.012025	0.007692	0.169293	0.011532	0.178490
	KSP	**0.62147**	**0.000501**	**0.001772**	**0.044165**	0.003312	0.169834
	CSD	0.62146	0.001267	0.003573	0.063101	0.003366	0.170954
	CSD-iPOT	0.61973	0.000617	0.002269	0.049259	**0.002470**	**0.169168**
CRPS	Non-calibrated	0.59211	0.085305	0.186298	0.457565	0.208641	0.387240
	KSP	**0.60051**	**0.000786**	**0.003384**	**0.057540**	**0.004217**	**0.175699**
	CSD	0.59756	0.008325	0.009216	0.131548	0.019657	0.222708
	CSD-iPOT	0.59810	0.002275	0.004011	0.085995	0.013559	0.185530
DeepHit	Non-calibrated	**0.63846**	0.000539	**0.001943**	**0.044299**	0.001352	0.162515
	KSP	0.63823	0.000541	0.001957	0.048270	0.001197	**0.162350**
	CSD	0.63793	0.000566	0.002485	0.049948	**0.001040**	0.162360
	CSD-iPOT	0.63746	**0.000518**	0.002548	0.049504	0.001268	0.162362
AFT	Non-calibrated	0.62854	0.170157	0.157761	0.699554	0.150305	0.323121
	KSP	0.62854	0.000696	0.003680	0.055747	0.019585	0.195871
	CSD	0.62854	**0.000492**	**0.002619**	**0.047241**	**0.001815**	**0.176144**
	CSD-iPOT	0.62854	0.001917	0.004316	0.085610	0.033533	0.203033

Table 14: Comparison of post-processing methods for the GBSG dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.67055	0.000153	0.001249	0.031010	0.000037	**0.177364**
	KSP	0.67055	**0.000131**	**0.001108**	**0.027683**	0.000246	0.177451
	CSD	0.67055	0.001139	0.002872	0.064236	0.001849	0.180139
	CSD-iPOT	0.67055	0.000137	0.001822	0.029647	**0.000152**	0.177440
MTLR	Non-calibrated	**0.66518**	0.015595	0.047053	0.305066	0.027289	0.212137
	KSP	0.66422	0.008840	0.046362	0.161603	0.015227	0.206787
	CSD	0.66511	**0.000277**	**0.002028**	**0.038923**	**0.000476**	**0.184225**
	CSD-iPOT	0.66508	0.002382	0.008590	0.112182	0.002627	0.186904
Parametric	Non-calibrated	0.64735	0.004308	0.003882	0.097194	0.006469	0.189363
	KSP	0.65704	**0.000404**	**0.001651**	**0.040159**	**0.000619**	**0.182844**
	CSD	**0.65748**	0.000673	0.002655	0.046418	0.001565	0.184507
	CSD-iPOT	0.65686	0.000528	0.002326	0.042402	0.000863	0.183205
CRPS	Non-calibrated	0.59357	0.060048	0.128590	0.389503	0.103073	0.299431
	KSP	0.60163	**0.000383**	**0.001487**	**0.041330**	**0.000531**	**0.194706**
	CSD	0.60237	0.004895	0.006940	0.109091	0.007319	0.227361
	CSD-iPOT	**0.60250**	0.000579	0.002102	0.046562	0.001127	0.195382
DeepHit	Non-calibrated	0.66395	0.013217	0.042233	0.286991	0.022726	0.207548
	KSP	0.66380	0.009500	0.047356	0.167647	0.016446	0.208659
	CSD	0.66390	**0.000286**	**0.002090**	**0.038490**	**0.000545**	**0.184770**
	CSD-iPOT	**0.66401**	0.002268	0.008670	0.110511	0.002322	0.186840
AFT	Non-calibrated	0.66068	0.162642	0.139443	0.694223	0.178947	0.372955
	KSP	0.66068	0.000388	**0.001406**	**0.039076**	0.001558	**0.185855**
	CSD	0.66068	**0.000316**	0.001880	0.039659	**0.000557**	0.201437
	CSD-iPOT	0.66068	0.000733	0.002273	0.051802	0.005238	0.186443

Table 15: Comparison of post-processing methods for the NACD dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.74482	**0.002403**	**0.003013**	**0.069770**	0.002771	0.148203
	KSP	0.74482	0.003181	0.003339	0.071627	0.002512	**0.148069**
	CSD	**0.74483**	0.006150	0.006106	0.109203	0.002737	0.150490
	CSD-iPOT	0.74481	0.002585	0.003708	0.069948	**0.002100**	0.148205
MTLR	Non-calibrated	**0.75232**	0.002727	0.004358	0.078424	0.008610	0.158760
	KSP	**0.75232**	**0.000394**	0.002154	**0.040311**	0.003399	0.151300
	CSD	0.75155	0.000434	**0.002150**	0.043359	**0.002094**	**0.150096**
	CSD-iPOT	0.75168	0.000470	0.002770	0.043038	0.004965	0.153054
Parametric	Non-calibrated	0.73739	0.002980	0.007046	0.091661	0.000875	0.151496
	KSP	0.73811	**0.000413**	**0.001920**	**0.040499**	0.002716	**0.149560**
	CSD	**0.73877**	0.001803	0.003523	0.070777	**0.000750**	0.149815
	CSD-iPOT	0.73868	0.000715	0.003684	0.056430	0.000980	0.149651
CRPS	Non-calibrated	0.71759	0.043051	0.063536	0.301582	0.089002	0.250414
	KSP	**0.73131**	**0.000485**	**0.001559**	**0.041403**	**0.006499**	**0.156702**
	CSD	0.72759	0.009327	0.009646	0.142307	0.025258	0.187515
	CSD-iPOT	0.72959	0.001852	0.003063	0.071834	0.013876	0.164684
DeepHit	Non-calibrated	**0.73923**	0.001121	0.003564	0.059614	0.003978	0.156312
	KSP	0.73889	**0.000408**	**0.001647**	**0.039227**	0.003167	0.152809
	CSD	0.73909	0.000612	0.002525	0.051245	**0.002205**	**0.150479**
	CSD-iPOT	0.73895	0.000484	0.002530	0.043515	0.003355	0.154051
AFT	Non-calibrated	0.75251	0.163167	0.126591	0.658877	0.080887	0.248429
	KSP	0.75251	**0.000341**	**0.001746**	**0.038208**	0.006366	**0.149097**
	CSD	0.75251	0.000373	0.002111	0.040763	**0.005790**	0.173913
	CSD-iPOT	0.75251	0.000640	0.002636	0.048710	0.009616	0.151260

Table 16: Comparison of post-processing methods for the NB-SEQ dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.74406	0.000353	0.001233	0.036594	0.000117	0.048694
	KSP	0.74406	**0.000124**	**0.001029**	**0.024159**	**0.000114**	**0.048526**
	CSD	0.74406	0.000817	0.004176	0.058940	0.000592	0.049299
	CSD-iPOT	0.74406	0.000244	0.004492	0.053483	0.000413	0.048943
MTLR	Non-calibrated	0.67586	**0.000151**	0.000985	**0.026283**	0.000248	**0.056395**
	KSP	**0.67598**	0.000201	**0.000969**	0.028721	**0.000229**	0.056408
	CSD	0.67564	0.000592	0.005005	0.058357	0.000697	0.057094
	CSD-iPOT	0.67570	0.000283	0.005322	0.055453	0.000551	0.056798
Parametric	Non-calibrated	0.71868	0.020123	0.054895	0.217614	0.039669	0.088938
	KSP	0.71868	**0.000194**	**0.000847**	**0.028462**	**0.011985**	**0.057502**
	CSD	0.71868	0.010833	0.008192	0.155574	0.012509	0.066510
	CSD-iPOT	0.71868	0.001833	0.005667	0.082260	0.018628	0.059964
CRPS	Non-calibrated	0.74602	0.049548	0.082799	0.399175	0.075753	0.130153
	KSP	0.74602	**0.000289**	**0.001012**	**0.033242**	**0.001293**	**0.049608**
	CSD	0.74602	0.002066	0.004542	0.079292	0.013118	0.071306
	CSD-iPOT	0.74602	0.005656	0.009961	0.140358	0.013339	0.059066
DeepHit	Non-calibrated	0.74238	**0.000169**	**0.000836**	**0.027303**	0.000137	0.047711
	KSP	0.74243	0.000196	0.000896	0.028329	**0.000125**	**0.047706**
	CSD	0.74267	0.000534	0.004087	0.061122	0.000552	0.048342
	CSD-iPOT	**0.74273**	0.000275	0.004866	0.057137	0.000414	0.048151
AFT	Non-calibrated	0.45964	0.195759	0.279953	0.750892	0.177522	0.249208
	KSP	0.45964	0.001428	0.005948	**0.068017**	0.040891	0.134525
	CSD	0.45964	**0.001277**	0.006713	0.069315	**0.001830**	**0.071353**
	CSD-iPOT	0.45964	0.001775	**0.005227**	0.080487	0.048066	0.128660

Table 17: Comparison of post-processing methods for the SUPPORT dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.61233	**0.000055**	**0.000474**	0.019310	**0.000017**	0.191039
	KSP	0.61233	0.000057	0.000475	**0.018416**	0.000057	**0.191034**
	CSD	0.61233	0.000336	0.001659	0.036604	0.000435	0.191591
	CSD-iPOT	0.61233	0.000057	0.001501	0.027010	0.000129	0.191146
MTLR	Non-calibrated	0.57371	0.000834	0.003760	0.061424	0.012056	0.220228
	KSP	0.57312	0.000490	0.002672	0.040363	0.007750	0.215751
	CSD	0.57336	**0.000163**	**0.002190**	**0.031640**	**0.000372**	**0.208278**
	CSD-iPOT	**0.57380**	0.000389	0.003534	0.039351	0.007746	0.215638
Parametric	Non-calibrated	0.59260	0.002252	0.003988	0.084550	0.005979	0.198560
	KSP	0.60439	0.000243	0.001349	0.034436	**0.000275**	**0.193467**
	CSD	**0.60485**	0.000440	0.001883	0.037064	0.004318	0.197069
	CSD-iPOT	0.60386	**0.000182**	**0.001047**	**0.025885**	0.001020	0.193504
CRPS	Non-calibrated	0.55840	0.141252	0.434019	0.658101	0.401180	0.615518
	KSP	0.57109	0.015244	0.209804	0.248033	0.011795	0.224882
	CSD	**0.57160**	**0.007093**	**0.017108**	**0.127136**	0.023034	0.256595
	CSD-iPOT	0.57113	0.025275	0.101719	0.380558	**0.003503**	**0.216977**
DeepHit	Non-calibrated	**0.60214**	0.000890	0.003056	0.057871	0.010317	0.206596
	KSP	0.60133	0.000356	**0.001975**	0.035817	0.007007	0.203254
	CSD	0.60162	**0.000192**	0.002228	**0.033087**	**0.001072**	**0.196510**
	CSD-iPOT	0.60138	0.000334	0.003204	0.035553	0.007413	0.202907
AFT	Non-calibrated	0.56930	0.150578	0.101386	0.641556	0.056436	0.268950
	KSP	0.56930	0.000226	0.001612	0.032202	**0.000245**	**0.207191**
	CSD	0.56930	**0.000175**	0.003102	0.041830	0.000897	0.214866
	CSD-iPOT	0.56930	0.000210	**0.000876**	**0.029300**	0.001178	0.208210

Table 18: Comparison of post-processing methods for the MIMIC-III dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.95839	0.036649	0.064547	0.250907	0.104279	0.276509
	KSP	0.95826	0.003305	**0.007044**	**0.091109**	0.071617	0.203039
	CSD	**0.95842**	**0.002873**	0.022981	0.113186	0.082850	0.222409
	CSD-iPOT	**0.95842**	0.003150	0.016172	0.099244	**0.069556**	**0.200904**
MTLR	Non-calibrated	**0.94494**	0.015402	0.035474	0.232722	0.000410	0.014391
	KSP	0.94218	**0.003581**	**0.004200**	0.084666	**0.000405**	**0.012505**
	CSD	0.94423	0.004565	0.007386	0.091212	0.000573	0.013220
	CSD-iPOT	0.94230	0.003792	0.005829	**0.084427**	0.000439	0.012597
Parametric	Non-calibrated	0.96204	0.042814	0.113358	0.339424	**0.000130**	0.012483
	KSP	0.96204	0.019627	**0.020349**	**0.170588**	0.000163	**0.006789**
	CSD	0.96204	0.020498	0.031328	0.243077	0.000174	0.010146
	CSD-iPOT	0.96204	**0.019147**	0.029920	0.173745	0.000178	0.006854
CRPS	Non-calibrated	0.95342	0.214466	0.258780	0.759872	0.020505	0.072193
	KSP	0.95339	0.013238	**0.012086**	0.149600	0.000378	0.012503
	CSD	**0.95398**	**0.001911**	0.013170	**0.111758**	0.009601	0.052380
	CSD-iPOT	0.95339	0.013594	0.017937	0.153465	**0.000357**	**0.012327**
DeepHit	Non-calibrated	0.96406	0.013780	**0.012673**	**0.157457**	**0.000064**	**0.005148**
	KSP	0.96425	0.021969	0.024353	0.174115	0.000084	0.005210
	CSD	**0.96437**	**0.021553**	0.038450	0.181557	0.000085	0.005288
	CSD-iPOT	0.96423	0.021580	0.037331	0.179908	0.000091	0.005276
AFT	Non-calibrated	0.88037	0.224182	0.416392	0.821818	0.019950	0.057193
	KSP	0.88037	0.005902	0.019573	0.125796	**0.000679**	0.021909
	CSD	0.88037	**0.000424**	0.009171	0.050827	0.002049	0.035906
	CSD-iPOT	0.88037	0.000573	**0.003317**	**0.046279**	0.000941	**0.020705**

Table 19: Comparison of post-processing methods for the SEER-liver dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.63374	0.000070	0.000569	0.022519	**0.000013**	**0.137621**
	KSP	0.63374	**0.000030**	**0.000178**	**0.012719**	0.000051	0.137654
	CSD	**0.63406**	0.000201	0.005539	0.053343	0.000521	0.138246
	CSD-iPOT	**0.63406**	0.000131	0.004846	0.049048	0.000371	0.138145
MTLR	Non-calibrated	0.63083	0.000188	**0.000557**	**0.036149**	0.002314	0.142314
	KSP	0.63077	0.000373	0.000820	0.039137	0.002094	0.142062
	CSD	0.63087	0.000169	0.005012	0.052282	**0.000762**	**0.140381**
	CSD-iPOT	**0.63088**	**0.000156**	0.004599	0.046668	0.002966	0.142248
Parametric	Non-calibrated	0.62927	0.001078	0.005251	0.056526	0.001723	0.137791
	KSP	0.63209	0.000167	**0.000810**	**0.023572**	**0.000895**	**0.137527**
	CSD	**0.63227**	0.000144	0.003663	0.039992	0.001917	0.138099
	CSD-iPOT	0.63208	**0.000143**	0.003052	0.034825	0.001836	0.138225
CRPS	Non-calibrated	0.60369	0.165436	0.455876	0.659311	0.277786	0.434392
	KSP	**0.60940**	0.001002	0.008353	0.056664	**0.000493**	0.147257
	CSD	0.60585	0.001337	0.003582	0.063189	0.014840	0.176125
	CSD-iPOT	0.60835	**0.000326**	**0.002620**	**0.037597**	0.001070	**0.145338**
DeepHit	Non-calibrated	0.63377	0.000210	0.000613	**0.036041**	0.002363	0.142056
	KSP	0.63366	0.000438	**0.000832**	0.039767	0.002216	0.141942
	CSD	**0.63383**	0.000173	0.004892	0.053137	**0.000805**	**0.140133**
	CSD-iPOT	0.63364	**0.000167**	0.004800	0.047828	0.002833	0.141958
AFT	Non-calibrated	0.62661	0.002288	0.008219	0.081952	0.003858	0.141641
	KSP	0.62661	0.000460	0.002309	0.036500	0.003495	0.142036
	CSD	**0.62691**	**0.000119**	0.003200	0.039183	**0.002418**	**0.140872**
	CSD-iPOT	**0.62691**	0.000120	**0.001753**	**0.026187**	0.003806	0.141918

Table 20: Comparison of post-processing methods for the SEER-stomach dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.70116	0.000180	0.000908	0.027420	**0.000017**	0.138086
	KSP	0.70116	**0.000020**	**0.000190**	**0.010721**	0.000096	**0.137950**
	CSD	**0.70134**	0.000790	0.005599	0.052845	0.000740	0.139080
	CSD-iPOT	**0.70134**	0.000120	0.004336	0.044332	0.000230	0.138501
MTLR	Non-calibrated	0.69538	0.000247	**0.000548**	**0.036700**	0.002508	0.147517
	KSP	0.69529	0.000480	0.000860	0.040512	0.002250	0.147458
	CSD	**0.69543**	0.000195	0.004382	0.050593	**0.000739**	**0.145290**
	CSD-iPOT	0.69532	**0.000161**	0.004533	0.048057	0.002610	0.147009
Parametric	Non-calibrated	0.69706	0.000673	0.003591	0.046981	0.001048	0.137726
	KSP	**0.70141**	0.000161	**0.000725**	**0.023063**	**0.001019**	**0.137680**
	CSD	0.70097	0.000867	0.003156	0.046788	0.001731	0.138664
	CSD-iPOT	0.70094	**0.000110**	0.001523	0.031439	0.001431	0.138318
CRPS	Non-calibrated	0.67345	0.146871	0.398096	0.615755	0.271412	0.440772
	KSP	**0.67678**	0.002785	0.018437	0.087667	0.003845	0.161579
	CSD	0.67524	0.005308	0.007148	0.109374	0.016138	0.176342
	CSD-iPOT	0.67611	**0.000700**	**0.004436**	**0.063063**	**0.001403**	**0.159007**
DeepHit	Non-calibrated	0.70345	0.000286	**0.000591**	**0.036736**	0.002516	0.144300
	KSP	0.70331	0.000479	0.000787	0.041623	0.002041	0.144021
	CSD	**0.70349**	0.000207	0.004104	0.048694	**0.000907**	**0.142015**
	CSD-iPOT	0.70327	**0.000151**	0.004071	0.046705	0.002762	0.143716
AFT	Non-calibrated	0.68993	0.001444	0.005714	0.064478	**0.001918**	**0.144443**
	KSP	0.68993	0.000211	**0.001181**	**0.025636**	0.002197	0.144783
	CSD	**0.69010**	0.000216	0.003218	0.036171	0.002119	0.144571
	CSD-iPOT	**0.69010**	**0.000098**	0.001515	0.028447	0.002509	0.145438

Table 21: Comparison of post-processing methods for the SEER-lung dataset. Higher C-index values indicate better discrimination, while lower values are preferred for all other calibration metrics.

Model	Method	C-index$^{\uparrow}$	S-cal(20)$^{\downarrow}$	D-cal(20)$^{\downarrow}$	KS-cal$^{\downarrow}$	KM-cal$^{\downarrow}$	IBS$^{\downarrow}$
DeepSurv	Non-calibrated	0.68068	0.000254	0.001268	0.032123	**0.000012**	0.106781
	KSP	0.68068	**0.000032**	**0.000129**	**0.011249**	0.000162	**0.106655**
	CSD	**0.68087**	0.000211	0.006196	0.056819	0.000573	0.107374
	CSD-iPOT	**0.68087**	0.000139	0.005479	0.050332	0.000455	0.107197
MTLR	Non-calibrated	0.67676	**0.000083**	**0.000531**	0.044664	**0.000273**	**0.109589**
	KSP	0.67671	0.000307	0.000624	0.044878	0.000338	0.109683
	CSD	0.67695	0.000127	0.004891	0.051604	0.000394	0.109700
	CSD-iPOT	**0.67699**	0.000119	0.004575	**0.044352**	0.000584	0.110014
Parametric	Non-calibrated	0.67932	0.000732	0.004613	0.051627	0.000897	**0.106561**
	KSP	0.68106	0.000243	**0.001297**	**0.026156**	**0.000780**	0.106744
	CSD	**0.68114**	0.000179	0.004534	0.047126	0.001264	0.107087
	CSD-iPOT	0.68096	**0.000095**	0.002838	0.044666	0.001082	0.107186
CRPS	Non-calibrated	0.65632	0.192684	0.499422	0.692678	0.287116	0.419652
	KSP	**0.66622**	0.000913	0.013576	0.061543	0.001476	0.117411
	CSD	0.65819	0.001192	0.004206	0.069645	0.014257	0.148662
	CSD-iPOT	0.66566	**0.000407**	**0.004188**	**0.052539**	**0.000993**	**0.114890**
DeepHit	Non-calibrated	0.68155	**0.000093**	**0.000546**	0.044664	**0.000268**	**0.108290**
	KSP	0.68151	0.000305	0.000648	0.044670	0.000327	0.108375
	CSD	0.68214	0.000127	0.004850	0.051287	0.000388	0.108417
	CSD-iPOT	**0.68218**	0.000118	0.004547	**0.043773**	0.000590	0.108751
AFT	Non-calibrated	0.67360	0.002282	0.007794	0.077526	0.001585	**0.111226**
	KSP	0.67360	0.000331	**0.001690**	**0.029455**	0.001573	0.111337
	CSD	**0.67378**	0.000136	0.005106	0.047482	**0.001409**	0.111230
	CSD-iPOT	**0.67378**	**0.000080**	0.002245	0.041804	0.001808	0.111826

G Calibration plot

In this section, we present calibration plots across datasets and models.

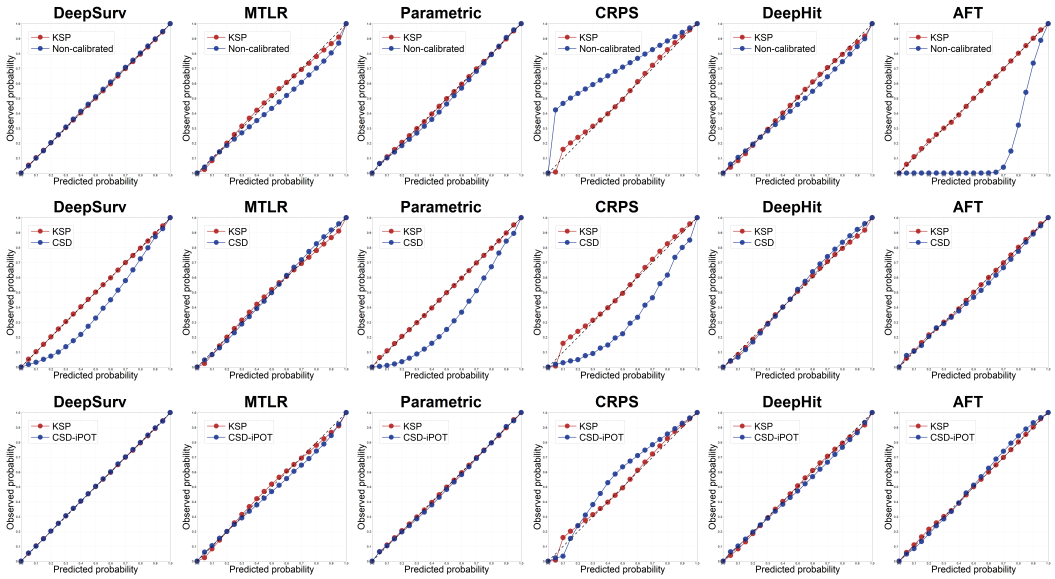


Figure 15: Calibration plot across six models for the WHAS dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

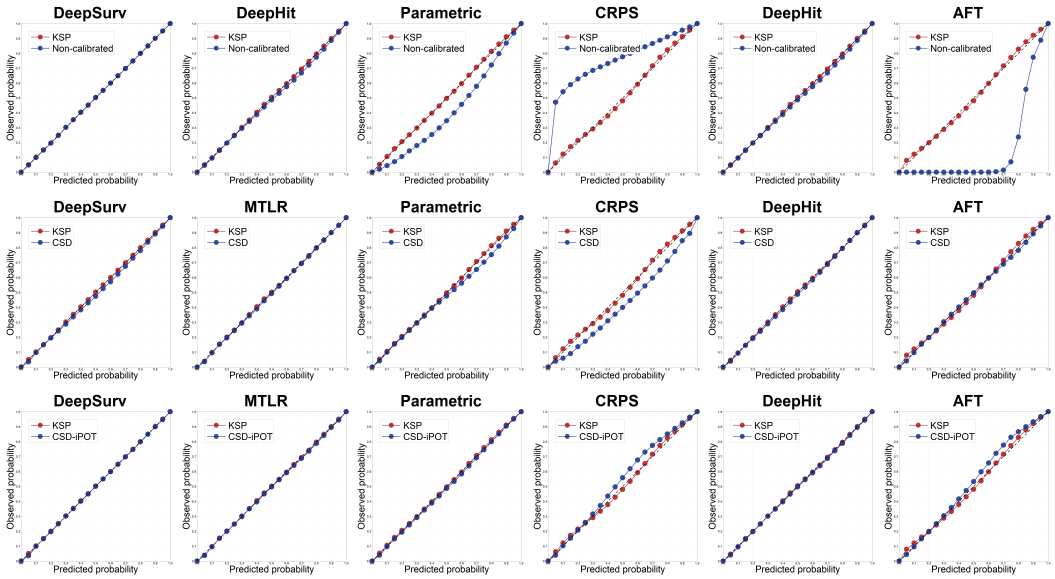


Figure 16: Calibration plot across six models for the METABRIC dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

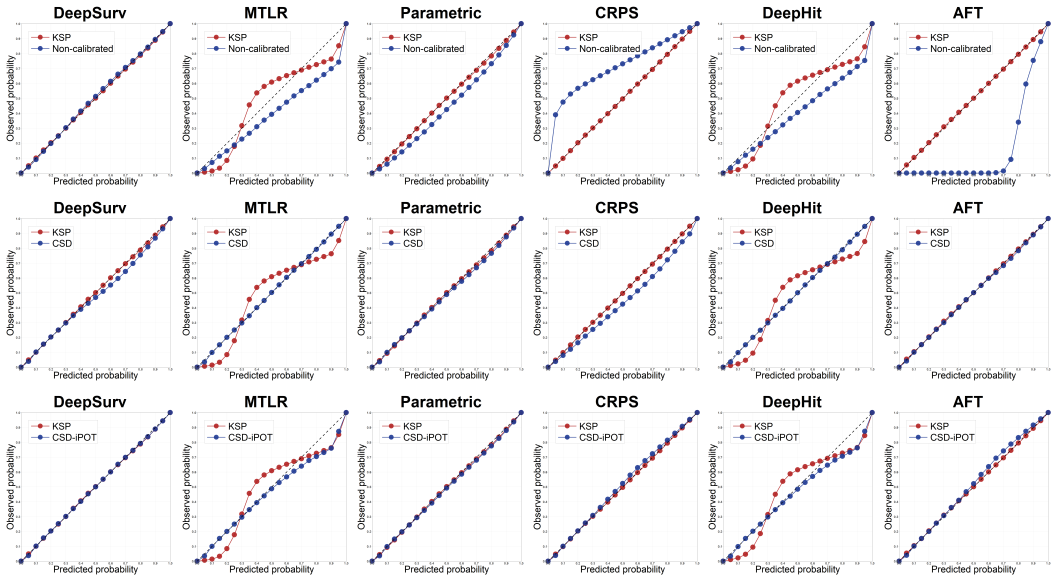


Figure 17: Calibration plot across six models for the GBSG dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

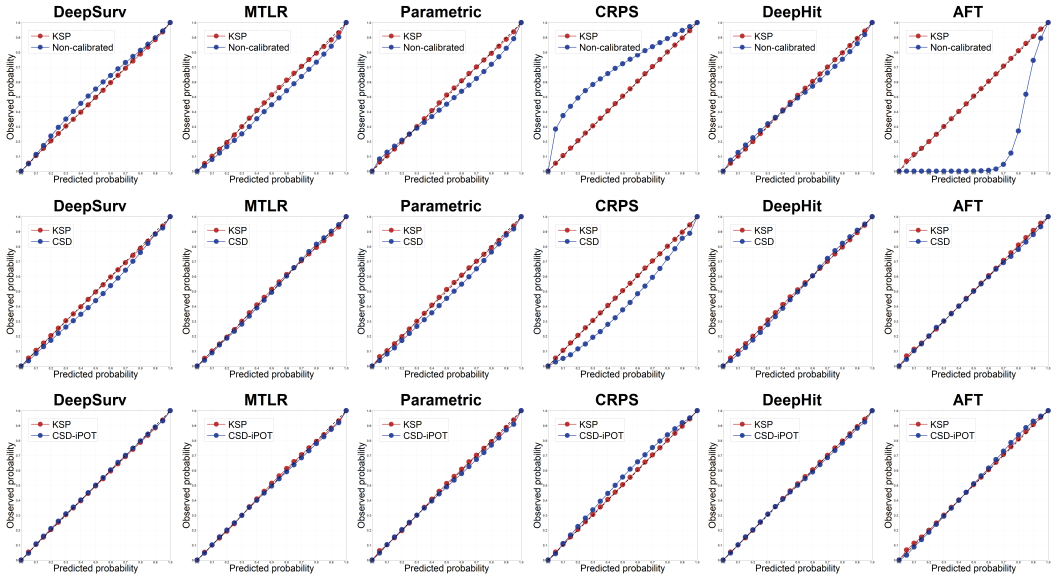


Figure 18: Calibration plot across six models for the NACD dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

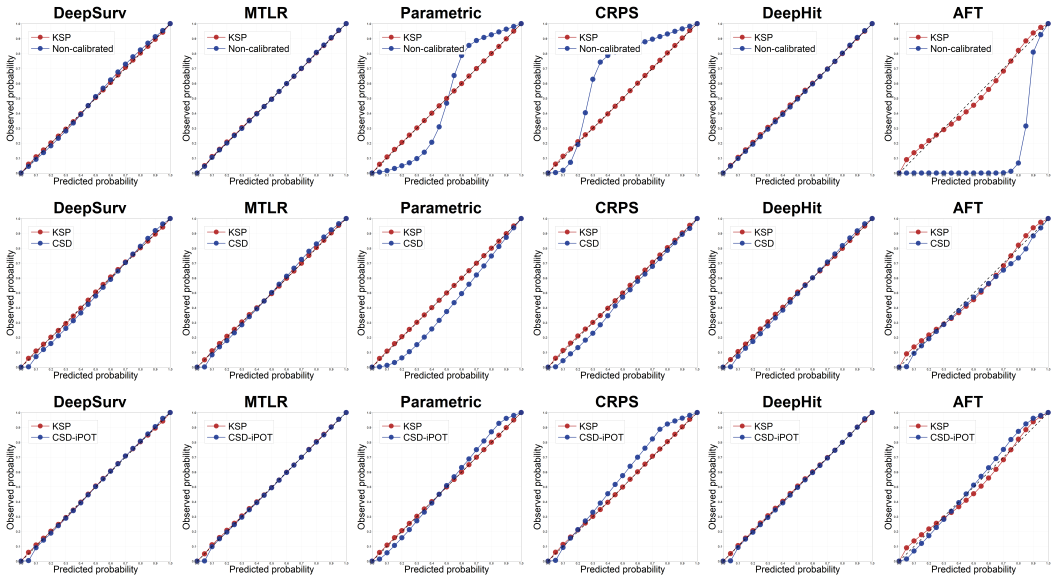


Figure 19: Calibration plot across six models for the NB-SEQ dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

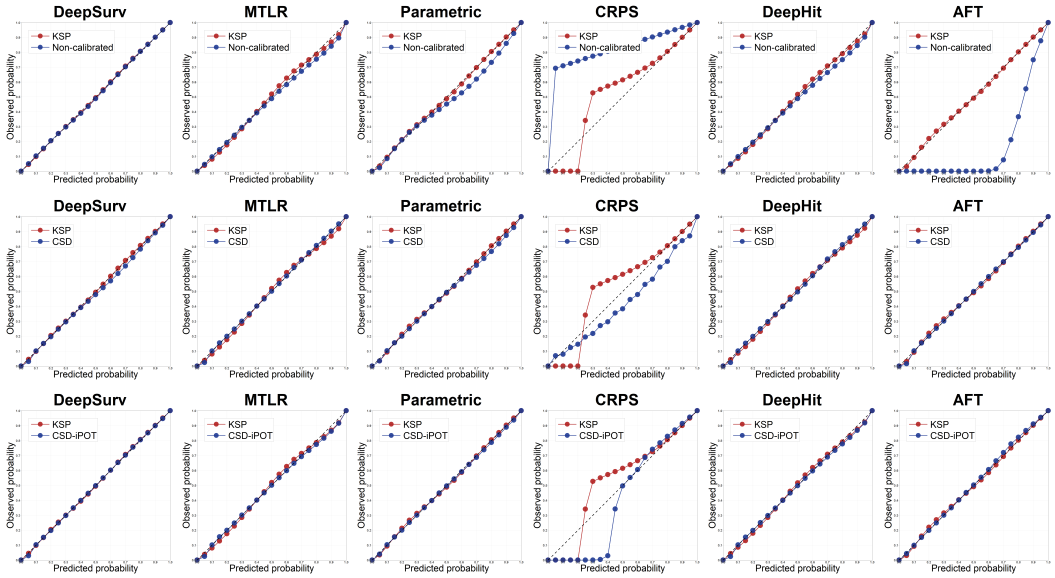


Figure 20: Calibration plot across six models for the SUPPORT dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

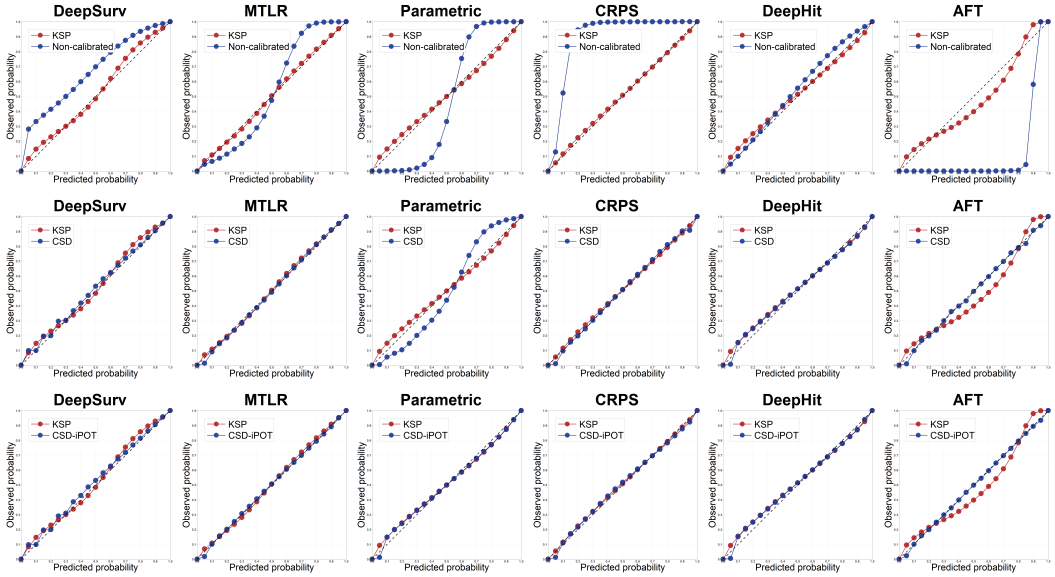


Figure 21: Calibration plot across six models for the MIMIC-III dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

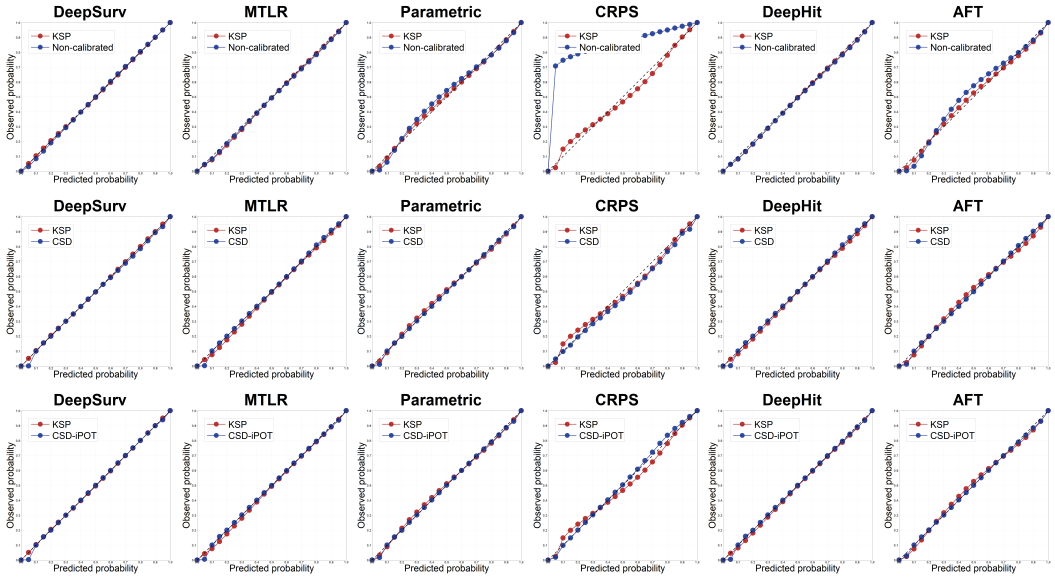


Figure 22: Calibration plot across six models for the SEER-liver dataset. Each model is evaluated under four calibration methods: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

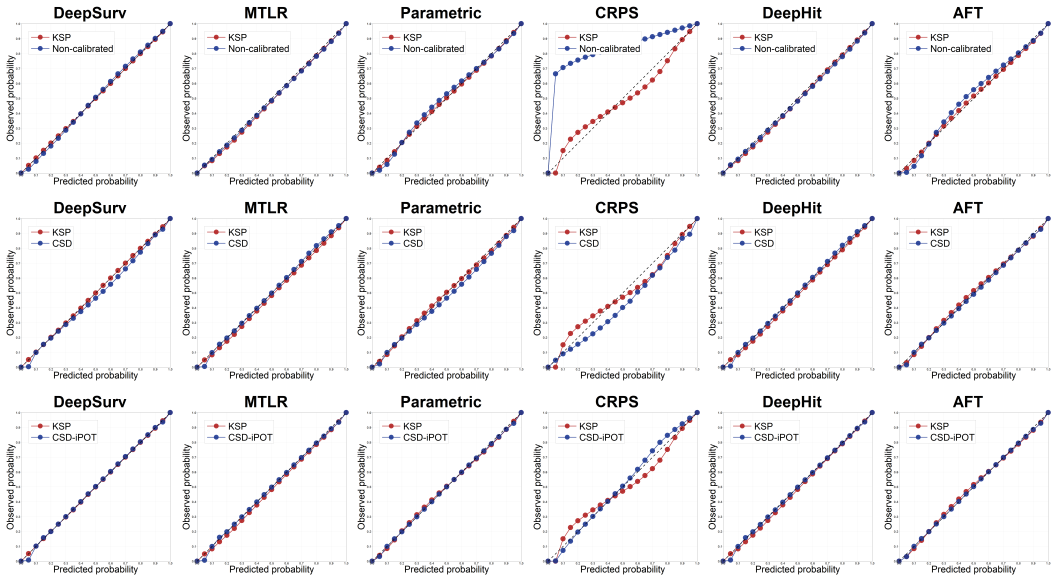


Figure 23: Calibration plot across six models for the SEER-stomach dataset. Each model is evaluated under four calibration settings: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

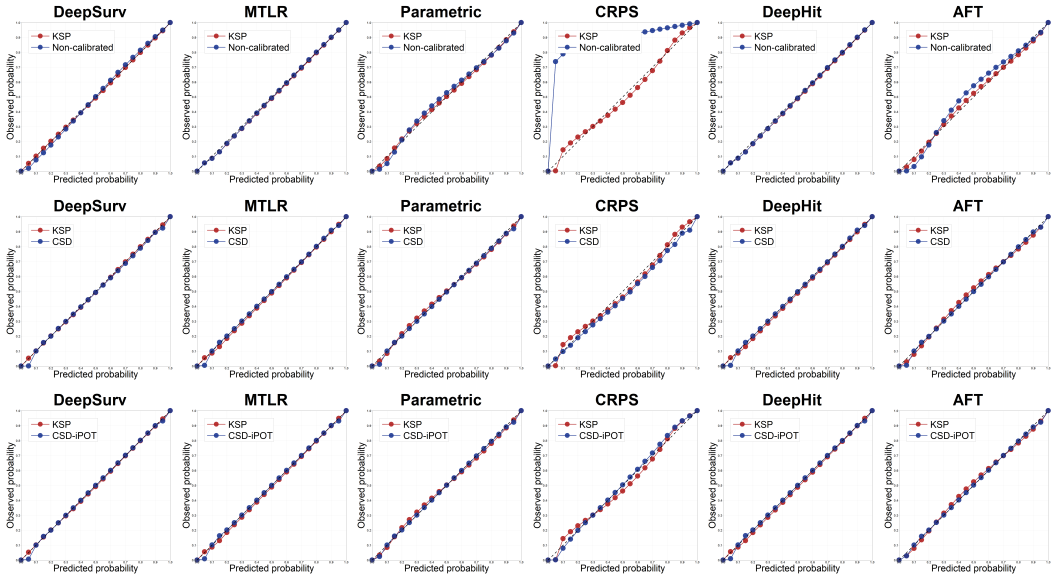


Figure 24: Calibration plot across six models for the SEER-lung dataset. Each model is evaluated under four calibration settings: Non-calibrated, KSP, CSD, and CSD-iPOT. A calibration slope close to 1 indicates better agreement between predicted and observed probabilities.

H Evaluation of calibration errors with 10-bin

In this section, we compare the performance of S-cal and D-cal using evaluations with 10 bins. To determine bin locations, we consider two strategies: uniformly spaced bins and a scheme that allocates denser bins in the tails. The results are reported in a format similar to Table 1 in the main manuscript. For uniformly spaced bins, we use knot points at $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, while for tail-focused bins, we use $\{0.05, 0.10, 0.20, 0.40, 0.50, 0.60, 0.80, 0.90, 0.95\}$. We denote the former case as S-cal(10)^1, D-cal(10)^1, and the latter case as S-cal(10)^2, D-cal(10)^2.

When comparing non-calibrated baselines, KSP consistently outperforms them, regardless of the metric. For CSD, evaluations with S-cal(10)^1 and D-cal(10)^1 slightly favor CSD over KSP, whereas CSD-iPOT outperforms when evaluated using the same metrics. This is natural, since these methods adjust quantile levels exactly at the knot points of S-cal(10)^1 and D-cal(10)^1. However, when calibration errors are measured with denser bins in the tails, KSP again outperforms all methods, consistent with the results from S-cal(20), D-cal(20), and KS-cal. From this perspective, metrics that depend on the choice of intervals are neither consistent nor fair, highlighting the advantage of bin-free metrics for robust comparison.

Table 22: Summary of pairwise comparisons between post-processing methods. The table shows the number of cases where KSP outperforms its counterpart, is outperformed, or yields a tie. Numbers in parentheses indicate statistically significant differences based on a one-sided t-test at the 0.05 level.

Method	S-cal(10)^1	D-cal(10)^1	S-cal(10)^2	D-cal(10)^2	S-cal(20)	D-cal(20)	KS-cal
KSP	**48 (45)**	**46 (43)**	**47 (44)**	**45 (42)**	**46 (45)**	**46 (43)**	**47 (45)**
Non-calibrated	12 (8)	14 (11)	13 (7)	15 (9)	13 (7)	14 (6)	13 (5)
Ties	1	0	0	0	1	0	0
KSP	**36 (28)**	**32 (29)**	**37 (28)**	**41 (37)**	**36 (29)**	**48 (45)**	**51 (42)**
CSD	24 (19)	28 (26)	23 (15)	19 (17)	24 (19)	12 (10)	9 (8)
Ties	0	0	0	0	0	0	0
KSP	26 (14)	25 (14)	**36 (20)**	**44 (33)**	**32 (21)**	**46 (39)**	**44 (29)**
CSD-iPOT	**34 (23)**	**35 (27)**	24 (13)	16 (15)	28 (19)	14 (13)	16 (11)
Ties	0	0	0	0	0	0	0

I Ablation study on KSP

In this section, we present an ablation study to investigate the effectiveness of KSP. For simplicity, we conduct experiments using the MIMIC-III dataset with three models: DeepSurv, MTLR, and CRPS. This dataset is characterized by a relatively high calibration error, making it a suitable benchmark for evaluating post-processing methods.

I.1 Ablation study 1: link function

We evaluate KSP using various link functions, including the logit, inverse hyperbolic tangent (atanh), three inverse CDFs with range $(-\infty, \infty)$—probit, inverse Cauchy, and inverse Laplace—and the complementary loglog (cloglog) function. Since atanh has a domain of $(-1, 1)$, we rescale the CDFs using $2\hat{F}_{\boldsymbol{\theta}} - 1$ for compatibility.

Table 23 summarizes the performance across DeepSurv, MTLR, and CRPS. While the logit function does not always yield the best score for every metric, it consistently delivers near-optimal performance across both discrimination and calibration. Atanh and probit show reasonable performance but are prone to numerical instability. The inverse Cauchy and inverse Laplace functions exhibit good performance on some metrics; however, they inherently assume specific parametric forms, making them less suitable for general-purpose use. Cloglog, intended to emphasize tail behavior, results in moderate performance. Considering both empirical stability and robustness, we adopt the logit function as the default link function in KSP.

Table 23: Comparison of different link functions G used in KSP for DeepSurv, MTLR, and CRPS on the MIMIC-III dataset. Bold values indicate the best performance within each model.

Model	G	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepSurv	logit	**0.95826**	0.003305	**0.007044**	**0.091109**	0.071617	0.203039
	atanh	0.94989	0.003882	0.022601	0.099237	0.071447	0.201032
	probit	0.93952	0.005747	0.046071	0.120879	0.076192	0.203363
	inverse Cauchy	0.95803	**0.003273**	0.009515	0.096415	**0.069989**	**0.200894**
	inverse Laplace	0.95804	0.004483	0.014649	0.111702	0.074329	0.208606
	cloglog	0.95797	0.004379	0.014051	0.109403	0.074808	0.209548
MTLR	logit	0.94089	**0.003552**	0.004132	**0.084090**	0.000409	0.012512
	atanh	0.94086	0.003632	**0.004106**	0.084987	0.000408	0.012503
	probit	**0.94155**	0.004588	0.011219	0.109099	0.000405	**0.012455**
	inverse Cauchy	0.94150	0.004397	0.009183	0.103863	0.000417	0.012737
	inverse Laplace	0.94112	0.003970	0.006271	0.095429	**0.000397**	0.012535
	cloglog	0.94090	0.003831	0.005342	0.091252	0.000409	0.012462
CRPS	logit	0.95339	0.013238	0.012086	0.149600	0.000378	0.012503
	atanh	0.95339	0.013067	**0.011765**	0.149861	0.000381	0.012514
	probit	0.95339	**0.012670**	0.012858	**0.145194**	0.000379	0.012507
	inverse Cauchy	**0.95341**	0.015722	0.027449	0.184878	0.000755	0.012856
	inverse Laplace	0.95339	0.013383	0.012086	0.158158	0.000399	**0.012402**
	cloglog	0.95339	0.013458	0.013700	0.149499	**0.000377**	0.012588

I.2 Ablation study 2: impact of using top-k deviations

The proposed KSP in the main text minimizes the KS-cal, which corresponds to the maximum deviation between the predicted and empirical CDFs. In this ablation study, we evaluate an alternative formulation that minimizes the sum of the k largest deviations instead of just the maximum alone. We report the performance of KSP under different values of k to examine whether aggregating multiple top deviations yields improved calibration or stability.

As shown in Table 24, increasing k leads to marginal improvements in calibration metrics; however, the gains are minimal. Notably, for MTLR, larger k values result in a decline in the C-index and a rise in calibration errors. These findings indicate that minimizing only the maximum deviation ($k = 1$) is sufficient to ensure effective calibration without sacrificing predictive performance.

Table 24: Ablation study on the choice of top-k deviations in the KSP. We report results for $k \in \{1, 5, 10, 50, 100\}$ across three models (DeepSurv, MTLR, CRPS) on MIMIC-III. The case $k = 1$ corresponds to the original KSP formulation using only the maximum deviation.

Model	Top k	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepSurv	1	0.95826	0.003305	0.007044	0.091109	0.071617	0.203039
	5	0.95826	0.003313	0.007051	0.090815	0.071621	0.203101
	10	0.95826	0.003252	**0.006966**	0.090368	0.071437	0.202842
	50	0.95826	0.003233	0.007143	0.090056	0.071202	**0.202644**
	100	0.95826	**0.003217**	0.007276	**0.089720**	**0.071107**	0.202672
MTLR	1	**0.94089**	**0.003552**	**0.004132**	**0.084090**	**0.000409**	0.012512
	5	0.94086	0.003703	0.004485	0.084696	0.000410	0.012518
	10	0.94082	0.003764	0.004679	0.085963	0.000411	**0.012511**
	50	0.94074	0.003840	0.005624	0.085756	0.000415	0.012516
	100	0.94078	0.003854	0.005538	0.085496	0.000416	0.012523
CRPS	1	0.95339	0.013238	0.012086	0.149600	0.000378	0.012503
	5	0.95339	0.013079	0.012191	**0.148102**	0.000373	0.012438
	10	0.95339	0.013475	0.012308	0.150593	0.000378	0.012480
	50	0.95339	0.013189	0.012114	0.150729	0.000372	0.012440
	100	0.95339	**0.012869**	**0.012076**	0.148483	**0.000370**	**0.012400**

I.3 Ablation study 3: Form of the deviation

Originally, the KSP was defined using the absolute value of deviations in Step 4. We also experiment with an alternative formulation using the squared error. Since the quadratic form penalizes large deviations more heavily than the absolute value, it may potentially lead to better calibration. Additionally, we report the optimization time in seconds (mean and standard deviation) to compare computational efficiency.

As shown in Table 25, although the squared objective slightly emphasizes larger deviations, the absolute form achieves nearly identical performance in both discrimination and calibration. More importantly, it is more efficient in training time.

Table 25: Effect of using squared vs. absolute deviation in KSP optimization. Time is reported in seconds (mean and standard deviation).

Model	Form	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS	Time
DeepSurv	Absolute	0.95826	**0.003305**	**0.007044**	0.091109	**0.071617**	**0.203039**	**3.0163 (0.4331)**
	Squared	0.95826	0.003318	0.007094	**0.090898**	0.071853	0.203535	3.4383 (0.6512)
MTLR	Absolute	0.94089	**0.003552**	**0.004132**	**0.084090**	0.000409	0.012512	**3.0810 (0.8526)**
	Squared	**0.94216**	0.003563	0.004232	0.084132	**0.000407**	**0.012502**	4.1825 (0.6761)
CRPS	Absolute	0.95339	0.013238	0.012086	0.149600	0.000378	0.012503	**3.4411 (0.8088)**
	Squared	0.95339	0.013238	0.012086	0.149600	0.000378	0.012503	6.6034 (1.8035)

I.4 Ablation study 4: hyperparameters

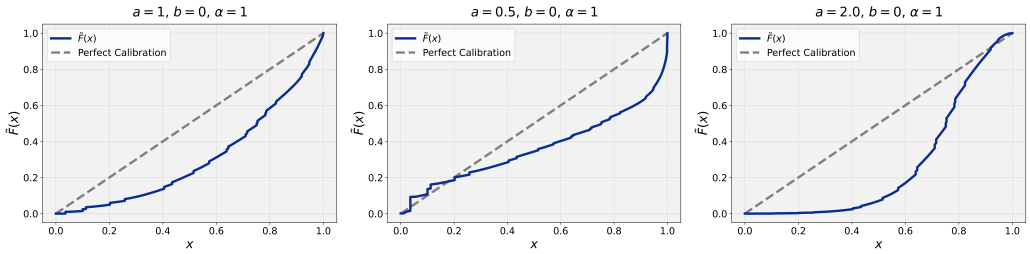
To assess the contribution of each hyperparameter used in KSP—namely a, b, and α—we conduct an ablation study by selectively choosing each component. This results in a total of 7 configurations, and the performance of each setting is reported to evaluate the individual and joint effects of the hyperparameters.

As shown in Table 26, using all three hyperparameters a, b, and α yields consistently strong performance across models. While the relative importance of each parameter may vary depending on the model, all three appear to contribute comparably to the overall calibration and discrimination quality.

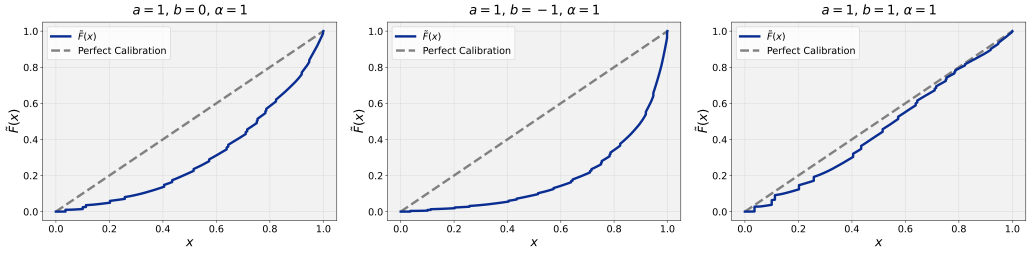
Table 26: Ablation study on the three hyperparameters a, b, and α used in KSP. Each row shows the performance when a subset of these hyperparameters is used.

Model	Hyperparameter	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepSurv	Non-calibrated	**0.95839**	0.036649	0.064547	0.250907	0.104279	0.276509
	a	0.95826	0.032545	0.050528	0.231501	0.102629	0.269723
	b	0.95826	0.011849	0.041974	0.183154	0.077582	0.230868
	α	0.95827	0.004572	0.016652	0.114728	0.074552	0.214151
	a, b	0.95826	0.005908	0.025273	0.127957	0.076690	0.213331
	a, α	0.95826	0.004852	0.017134	0.113908	0.075403	0.210912
	b, α	0.95826	0.004119	0.011969	0.105770	0.073772	0.207425
	a, b, α	0.95826	**0.003305**	**0.007044**	**0.091109**	**0.071617**	**0.203039**
MTLR	Non-calibrated	**0.94494**	0.015402	0.035474	0.232722	0.000410	0.014391
	a	0.94212	**0.002954**	0.010749	0.092681	0.000403	0.012762
	b	0.94075	0.016474	0.037806	0.209210	0.000380	0.013446
	α	0.94077	0.017945	0.041025	0.216149	0.000415	0.013717
	a, b	0.94168	0.004578	0.011120	0.106987	0.000403	**0.012481**
	a, α	0.94164	0.004438	0.010614	0.105226	0.000406	**0.012481**
	b, α	0.94175	0.011401	0.029399	0.184595	**0.000346**	0.012655
	a, b, α	0.94089	0.003552	**0.004132**	**0.084090**	0.000409	0.012512
CRPS	Non-calibrated	**0.95342**	0.214466	0.258780	0.759872	0.020505	0.072193
	a	**0.95342**	0.083571	0.942654	0.499678	0.194741	0.236811
	b	0.95341	0.029452	0.065234	0.278981	0.000986	0.014843
	α	**0.95342**	0.048102	0.153166	0.357446	0.002043	0.021537
	a, b	0.95339	**0.013067**	0.012301	0.151768	0.000396	**0.012490**
	a, α	0.95340	0.043578	0.125829	0.340774	0.000908	0.017905
	b, α	0.95340	0.022179	0.043438	0.242998	0.000822	0.013355
	a, b, α	0.95339	0.013238	**0.012086**	**0.149600**	**0.000378**	0.012503

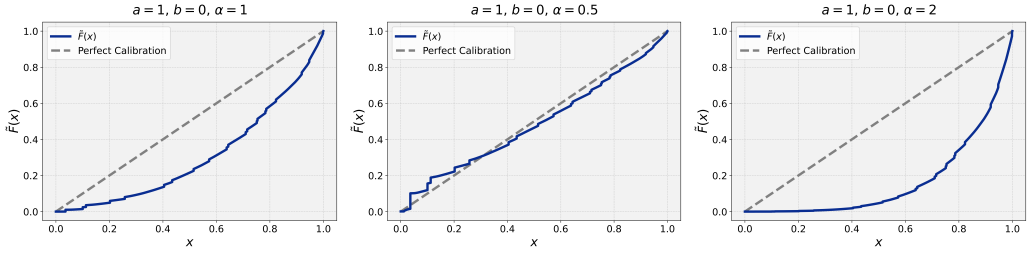
On the next page, Figure 25 illustrates how each hyperparameter affects $\tilde{F}(x)$.



(a) Effect of varying a with fixed $b = 0$, $\alpha = 1$



(b) Effect of varying b with fixed $a = 1$, $\alpha = 1$



(c) Effect of varying α with fixed $a = 1$, $b = 0$

Figure 25: Visualization of how each hyperparameter (a, b, α) influences $\tilde{F}(x)$. The leftmost plot in each row is the baseline: $a = 1$, $b = 0$, $\alpha = 1$.

J Adaptability of KS-cal to in-processing procedures

As discussed in the main paper, the original KS-cal can be adapted for in-processing by reformulating it as a penalty term:

$$\text{KS-cal}(k) = \sum_{j=1}^{k} \left(D_{(N-j+1)}\right)^2$$

Here, k determines how many of the largest deviations are penalized, controlling the trade-off between discrimination and calibration. Using only the maximum deviation ($k = 1$) can be unstable, as the evaluation point q_j varies during training with changes in $\hat{F}_{\boldsymbol{\theta}}$. To reduce this instability, we aggregate the top-k deviations. The squared form improves gradient behavior and allows for smoother optimization.

We evaluate $k \in \{1, 5, 10\}$ on the MIMIC-III dataset using DeepSurv, MTLR, and CRPS, and compare the results to X-cal with 20 bins, denoted as X-cal(20). To assess each method's ability to manage the trade-off between calibration and discrimination, we plot calibration error against the C-index for various values of λ and compute the area under the resulting curve. A smaller area indicates a better trade-off.

As shown in Figure 26, increasing λ generally reduces calibration error but also lowers the C-index. Larger values of k tend to yield more favorable trade-offs. While S-cal(20), D-cal(20), and KS-cal show consistent decreases with increasing λ, KM-cal and IBS do not, particularly for models other than DeepSurv.

Although KS-cal(10) clearly outperforms X-cal(20) in terms of area, the comparison may be misleading, as it includes points with low C-index values that are rarely acceptable in practice. Moreover, computing the full curve requires densely sampling λ, which incurs a high computational cost. Thus, we turn to selecting a single optimal λ per method.

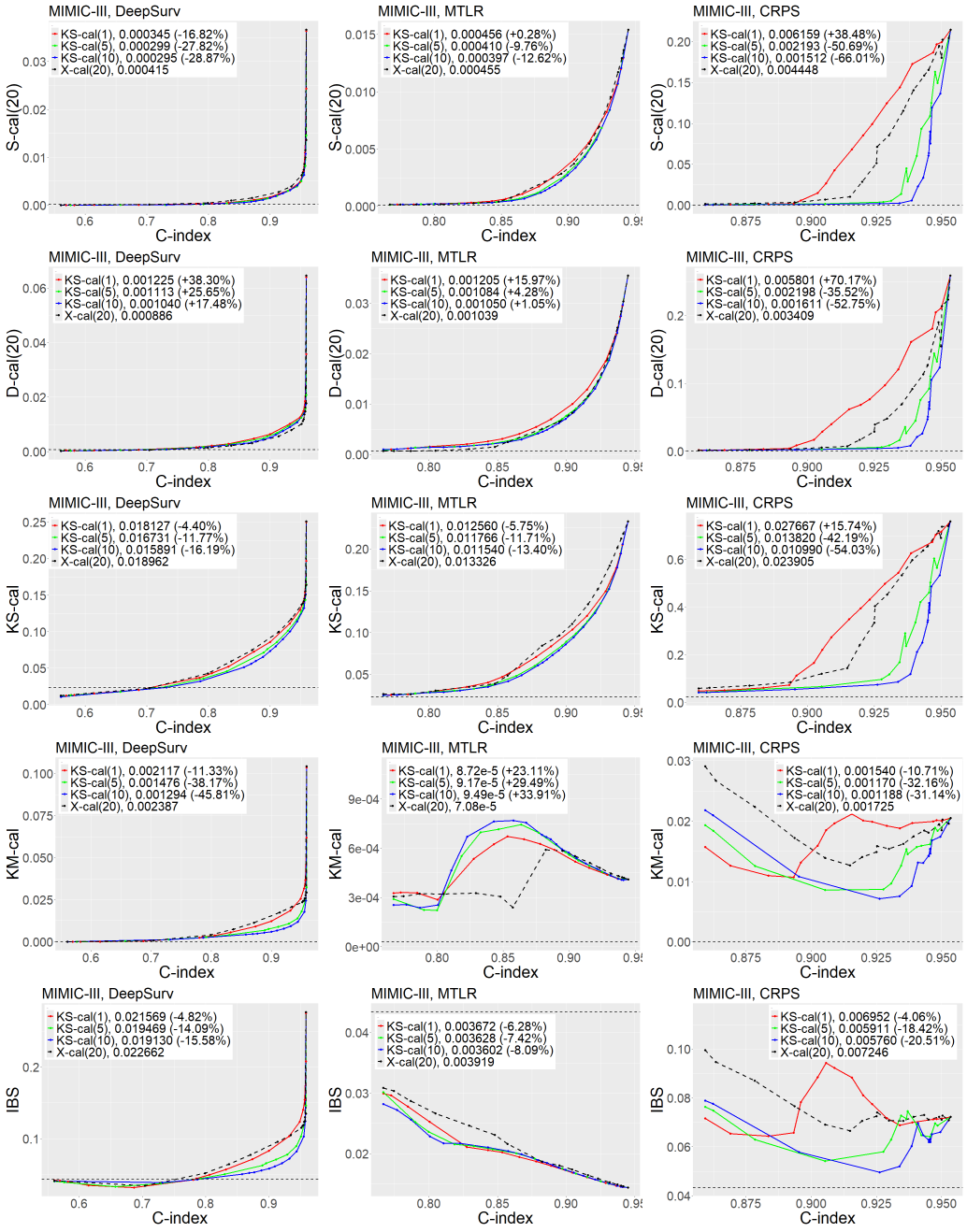


Figure 26: Trade-off between C-index and calibration errors on the MIMIC-III dataset using DeepSurv, MTLR, and CRPS. The legend indicates the area under the curve for each method, with the percentage decrease relative to X-cal shown in parentheses.

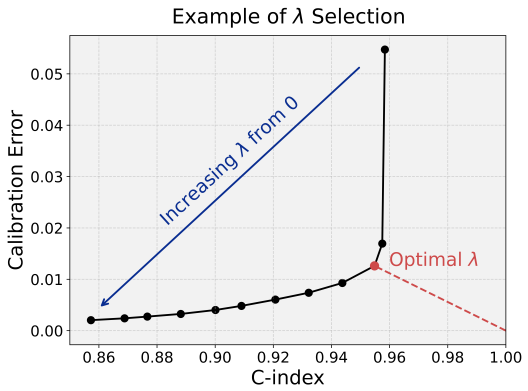


Figure 27: Trade-off between C-index and calibration error across different values of λ. The optimal point minimizes the Euclidean distance to $(1, 0)$.

Table 27: Example of finding optimal λ.

λ	C-index	S-cal(20)	D-cal(20)	KS-cal	CE	$1-$ C-index (scaled)	CE (scaled)	Distance	Optimal
0	0.95839	0.036661	0.064546	0.250918	0.054722	0.6135	1.0000	1.1732	X
0.5	0.95749	0.010718	0.017268	0.151156	0.016945	0.6268	0.2021	0.6586	O
1	0.95483	0.006342	0.014011	0.132187	0.012609	0.6660	0.1105	0.6751	X
2	0.94375	0.004108	0.010805	0.113729	0.009282	0.8294	0.0403	0.8304	X
3	0.93218	0.003108	0.009007	0.100064	0.007376	1.0000	0.0000	1	X

J.1 Selection of the regularization parameter

The core challenge in in-processing methods is choosing a regularization parameter λ that balances calibration and discrimination. While prior work often prioritizes one over the other (Karandikar et al., 2021), our approach seeks to balance both by minimizing the Euclidean distance to the ideal point $(1, 0)$ in the C-index–calibration error space:

$$\text{Distance} = \sqrt{(1 - \text{C-index})^2 + (\text{calibration error})^2}$$

We use S-cal(20), D-cal(20), and KS-cal to compute calibration error and exclude KM-cal and IBS due to their non-monotonic behavior with respect to λ. Since each metric shows different sensitivity to λ (Figure 26), we average the three and denote this composite as CE. We use the squared KS-cal in this average to account for its magnitude.

To ensure both components contribute equally to the distance, we apply the following min-max normalization:

$$\begin{aligned} 1 - \text{C-index} \quad &\Longrightarrow \quad \frac{1 - \text{C-index}}{1 - \min_\lambda (\text{C-index})} \\ \text{CE} \quad &\Longrightarrow \quad \frac{\text{CE} - \min_\lambda (\text{CE})}{\max_\lambda (\text{CE}) - \min_\lambda (\text{CE})} \end{aligned}$$

This normalization brings both metrics to the same scale before computing the distance. An example of the scaled values and computed distances is shown in Table 27.

Using this criterion, we compare the best-performing configurations of KS-cal(k), X-cal(20), and KSP. The results are summarized in Table 28. In-processing methods can sometimes yield lower calibration error than post-processing approaches, but this often comes at the cost of a reduced C-index. They may be appropriate when calibration is prioritized over predictive accuracy, despite a higher computational cost. Addressing this trade-off remains an important direction for future research.

Table 28: Comparison of in-processing methods with their optimal λ values and KSP on the MIMIC-III dataset using DeepSurv, MTLR, and CRPS.

Model	Method	C-index	S-cal(20)	D-cal(20)	KS-cal	KM-cal	IBS
DeepSurv	KS-cal(1)	0.93170	0.003199	0.010169	0.110807	0.018586	0.103577
	KS-cal(5)	0.92872	0.003000	0.009330	0.101107	0.010723	0.077831
	KS-cal(10)	0.93218	0.003108	0.009007	0.100064	0.009279	0.072327
	X-cal(20)	0.93422	0.003895	0.007950	0.117502	0.020680	0.104861
	KSP	0.95826	0.003305	0.007044	0.091109	0.071617	0.203039
MTLR	KS-cal(1)	0.88891	0.002555	0.007000	0.083445	0.000590	0.018006
	KS-cal(5)	0.90702	0.003735	0.009094	0.098823	0.000527	0.016721
	KS-cal(10)	0.90546	0.003365	0.008331	0.094365	0.000539	0.016840
	X-cal(20)	0.89491	0.002862	0.006477	0.096721	0.000586	0.018059
	KSP	0.94218	0.003581	0.004200	0.084666	0.000405	0.012505
CRPS	KS-cal(1)	0.91555	0.068442	0.062303	0.349527	0.021180	0.088294
	KS-cal(5)	0.93445	0.013617	0.016511	0.168024	0.012609	0.072749
	KS-cal(10)	0.93854	0.005743	0.008530	0.118517	0.009256	0.060236
	X-cal(20)	0.92507	0.051386	0.028927	0.335169	0.014853	0.072469
	KSP	0.95339	0.013238	0.012086	0.149600	0.000378	0.012503

K Additional information on experiment

Table 29 reports the post-processing time (in seconds) for KSP, CSD, and CSD-iPOT across all real datasets. Table 30 summarizes the hyperparameter settings used for training. All models were optimized using the ADAM optimizer with early stopping (patience = 200).

Table 29: Post-processing time (in seconds) based on KSP, CSD, and CSD-iPOT across datasets, reported as Mean (SD).

Dataset	Model	KSP	CSD	CSD-iPOT	Dataset	Model	KSP	CSD	CSD-iPOT
WHAS	DeepSurv	0.7678 (0.3811)	1.7969 (1.6016)	0.2228 (0.0076)	SUPPORT	DeepSurv	0.7058 (0.2725)	4.2659 (0.9457)	1.8757 (0.0417)
	MTLR	0.7283 (0.6594)	0.7995 (0.6289)	0.2590 (0.0137)		MTLR	1.5205 (0.6209)	3.2952 (0.4357)	2.0184 (0.0528)
	Parametric	2.2840 (1.0490)	0.5261 (0.3060)	0.2533 (0.0130)		Parametric	2.6104 (0.7900)	4.7179 (1.0102)	1.7611 (0.0585)
	CRPS	0.9598 (0.5243)	0.4331 (0.0076)	0.2282 (0.0069)		CRPS	1.4559 (0.9106)	2.8087 (0.0428)	1.7310 (0.0464)
	DeepHit	0.4653 (0.3115)	0.4900 (0.0101)	0.2544 (0.0075)		DeepHit	1.5589 (0.6242)	3.1376 (0.3221)	1.7854 (0.0279)
	AFT	1.0208 (0.1975)	15.5978 (1.2419)	0.2419 (0.0116)		AFT	0.7861 (0.2624)	30.0743 (2.5449)	1.6943 (0.0723)
METABRIC	DeepSurv	0.7950 (0.3403)	0.6054 (0.2809)	0.2940 (0.0110)	MIMIC-III	DeepSurv	3.0368 (0.4841)	2.4220 (0.4372)	1.6408 (0.0412)
	MTLR	1.1833 (0.5774)	0.6906 (0.2434)	0.2993 (0.0167)		MTLR	3.0571 (0.8218)	42.1341 (2.9090)	1.7573 (0.0561)
	Parametric	1.1773 (0.7385)	1.1316 (1.3924)	0.2933 (0.0114)		Parametric	1.6159 (0.5322)	34.6717 (2.7579)	1.7789 (0.0668)
	CRPS	0.8682 (0.4417)	1.1776 (1.1617)	0.2889 (0.0105)		CRPS	3.0967 (0.7620)	11.5253 (15.4780)	1.7573 (0.0517)
	DeepHit	1.0289 (0.5412)	0.5619 (0.1199)	0.2982 (0.0133)		DeepHit	0.9375 (0.4121)	9.6657 (7.2375)	1.7658 (0.0623)
	AFT	1.4309 (0.2378)	14.3036 (1.3326)	0.2627 (0.0090)		AFT	3.9551 (0.6115)	72.0553 (1.4179)	1.7570 (0.0672)
GBSG	DeepSurv	0.9000 (0.4123)	1.9335 (1.1640)	0.3427 (0.0125)	SEER-liver	DeepSurv	1.2076 (0.3057)	10.9716 (0.1832)	7.8048 (0.1496)
	MTLR	1.5787 (0.7238)	0.6346 (0.0108)	0.3503 (0.0126)		MTLR	0.3937 (0.1692)	12.7019 (0.1489)	9.5496 (0.0995)
	Parametric	1.0653 (0.5401)	0.6255 (0.0138)	0.3475 (0.0155)		Parametric	3.0441 (0.5889)	10.7083 (0.1670)	7.5654 (0.1371)
	CRPS	1.2178 (0.5923)	2.0031 (1.9268)	0.3455 (0.0192)		CRPS	3.4469 (0.7613)	10.7443 (0.1496)	7.4120 (0.1100)
	DeepHit	1.4526 (0.6882)	0.6363 (0.0272)	0.3456 (0.0043)		DeepHit	0.3226 (0.1697)	12.7441 (0.1791)	9.6160 (0.1950)
	AFT	1.1944 (0.1961)	16.0761 (0.8339)	0.3262 (0.0213)		AFT	3.2206 (0.8557)	22.6053 (2.7416)	7.0594 (0.1624)
NACD	DeepSurv	1.4460 (0.6432)	2.6621 (1.2473)	0.3734 (0.0099)	SEER-stomach	DeepSurv	1.0196 (0.2596)	17.1235 (2.0229)	10.8288 (0.1893)
	MTLR	1.5000 (0.6986)	0.7431 (0.1895)	0.3765 (0.0042)		MTLR	0.4129 (0.3515)	17.8684 (0.2084)	13.8391 (1.6858)
	Parametric	0.9123 (0.4075)	0.6825 (0.0500)	0.3766 (0.0111)		Parametric	3.1469 (0.8141)	15.0044 (1.2413)	10.5314 (0.3136)
	CRPS	1.4590 (0.5881)	0.6606 (0.0220)	0.3743 (0.0136)		CRPS	3.6969 (0.8608)	14.7305 (0.9472)	10.1087 (0.1355)
	DeepHit	0.9239 (0.5399)	1.0276 (0.7442)	0.3801 (0.0099)		DeepHit	0.4406 (0.3845)	17.8186 (0.2356)	13.6322 (0.3151)
	AFT	0.8895 (0.2228)	9.0016 (1.1539)	0.3283 (0.0104)		AFT	2.8866 (0.6546)	38.4584 (4.0847)	9.5965 (0.1246)
NB-SEQ	DeepSurv	0.6386 (0.2721)	1.4287 (0.0315)	0.8005 (0.0187)	SEER-lung	DeepSurv	2.5802 (0.5329)	209.7240 (4.3867)	194.2884 (4.3613)
	MTLR	0.6330 (0.2507)	1.5685 (0.4306)	0.8268 (0.0279)		MTLR	0.7715 (0.1797)	274.7291 (3.9333)	199.1179 (21.5456)
	Parametric	1.5743 (0.4695)	2.1082 (0.5367)	0.8223 (0.0384)		Parametric	6.4046 (1.6300)	243.2266 (32.0076)	194.5959 (26.4842)
	CRPS	2.1784 (0.4662)	19.0665 (3.1521)	0.8214 (0.0388)		CRPS	6.3627 (1.6321)	200.5405 (2.4764)	174.5297 (2.2676)
	DeepHit	0.7653 (0.3243)	1.4631 (0.0579)	0.8161 (0.0157)		DeepHit	0.3815 (0.2236)	270.7449 (27.0538)	207.6804 (12.4522)
	AFT	2.5767 (0.3960)	17.5392 (2.2009)	0.7266 (0.0251)		AFT	7.3328 (1.5618)	384.4848 (10.3814)	175.9503 (2.5391)

Table 30: Set-ups used in experiments.

Dataset	Model	Batch size	Learning rate	Maximum epoch	Dropout rate	Hidden layer	Number of bins
WHAS	DeepSurv	64	1e-4	1000	0.1	24/12/12	-
	MTLR	64	1e-3	2000	-	-	20
	Parametric	64	1e-3	1000	0.1	24/12/12	-
	CRPS	128	1e-3	1000	0.1	24/12/12	-
	DeepHit	64	1e-3	2000	0.1	24/12/12	40
	AFT	64	1e-1	1000	-	-	-
METABRIC	DeepSurv	128	1e-4	1000	0.3	36/18/18	-
	MTLR	64	1e-3	2000	-	-	20
	Parametric	128	1e-3	1000	0.3	36/18/18	-
	CRPS	256	1e-3	1000	0.3	36/18/18	-
	DeepHit	128	1e-3	2000	0.3	36/18/18	40
	AFT	128	1e-1	1000	-	-	-
GBSG	DeepSurv	128	1e-4	1000	0.3	28/14/14	-
	MTLR	128	1e-3	2000	-	-	20
	Parametric	128	1e-3	1000	0.3	28/14/14	-
	CRPS	256	1e-3	1000	0.3	28/14/14	-
	DeepHit	128	1e-3	2000	0.3	28/14/14	40
	AFT	128	1e-1	1000	-	-	-
NACD	DeepSurv	128	1e-4	1000	0.1	51/51/51	-
	MTLR	128	1e-3	2000	-	-	20
	Parametric	128	1e-3	1000	0.1	51/51/51	-
	CRPS	256	1e-3	1000	0.1	51/51/51	-
	DeepHit	128	1e-3	2000	0.1	51/51/51	40
	AFT	128	1e-1	1000	-	-	-
NB-SEQ	DeepSurv	256	1e-4	1000	0.3	96/48/48	-
	MTLR	128	1e-3	2000	-	-	20
	Parametric	256	1e-3	1000	0.3	96/48/48	-
	CRPS	512	1e-3	1000	0.3	96/48/48	-
	DeepHit	256	1e-3	2000	0.3	96/48/48	40
	AFT	256	1e-1	1000	-	-	-
SUPPORT	DeepSurv	512	1e-4	1000	0.3	56/28/28	-
	MTLR	512	1e-3	2000	-	-	20
	Parametric	512	1e-3	1000	0.3	56/28/28	-
	CRPS	1024	1e-2	1000	0.3	56/28/28	-
	DeepHit	512	1e-3	2000	0.3	56/28/28	40
	AFT	512	1e-1	1000	-	-	-
MIMIC-III	DeepSurv	512	1e-4	1000	0.1	60/30/30	-
	MTLR	512	1e-3	2000	-	-	20
	Parametric	512	1e-3	1000	0.1	60/30/30	-
	CRPS	1024	1e-3	1000	0.1	60/30/30	-
	DeepHit	512	1e-3	1000	0.1	60/30/30	100
	AFT	512	1e-1	1000	-	-	-
SEER-liver	DeepSurv	1000	1e-4	1000	0.1	60/30/30	-
	MTLR	1000	1e-2	2000	-	-	20
	Parametric	1000	1e-3	1000	0.1	60/30/30	-
	CRPS	2000	1e-3	1000	0.1	60/30/30	-
	DeepHit	1000	1e-3	2000	0.1	60/30/30	20
	AFT	1000	1e-1	1000	-	-	-
SEER-stomach	DeepSurv	1000	1e-4	1000	0.1	60/30/30	-
	MTLR	1000	1e-3	2000	-	-	10
	Parametric	1000	1e-3	1000	0.1	60/30/30	-
	CRPS	2000	1e-3	1000	0.1	60/30/30	-
	DeepHit	1000	1e-3	2000	0.1	60/30/30	10
	AFT	1000	1e-1	1000	-	-	-
SEER-lung	DeepSurv	5000	1e-4	1000	0.1	60/30/30	-
	MTLR	5000	1e-3	2000	-	-	10
	Parametric	5000	1e-3	1000	0.1	60/30/30	-
	CRPS	10000	1e-3	1000	0.1	60/30/30	-
	DeepHit	5000	1e-3	2000	0.1	60/30/30	10
	AFT	5000	1e-1	1000	-	-	-