

SurvUnc: A Meta-Model Based Uncertainty Quantification Framework for Survival Analysis

Yu Liu*
yu.liu@eng.ox.ac.uk
University of Oxford
Oxford, UK

Wei Yao Tao
twy030819@gmail.com
University of Oxford
Oxford, UK

Tong Xia
tx229@cam.ac.uk
University of Cambridge
Cambridge, UK

Simon Knight
simon.knight@nds.ox.ac.uk
University of Oxford
Oxford, UK

Tingting Zhu
tingting.zhu@eng.ox.ac.uk
University of Oxford
Oxford, UK

Abstract

Survival analysis, which estimates the probability of event occurrence over time from censored data, is fundamental in numerous real-world applications, particularly in high-stakes domains such as healthcare and risk assessment. Despite advances in numerous survival models, quantifying the uncertainty of predictions from these models remains underexplored and challenging. The lack of reliable uncertainty quantification limits the interpretability and trustworthiness of survival models, hindering their adoption in clinical decision-making and other sensitive applications. To bridge this gap, in this work, we introduce *SurvUnc*, a novel meta-model based framework for post-hoc uncertainty quantification for survival models. *SurvUnc* introduces an anchor-based learning strategy that integrates concordance knowledge into meta-model optimization, leveraging pairwise ranking performance to estimate uncertainty effectively. Notably, our framework is model-agnostic, ensuring compatibility with any survival model without requiring modifications to its architecture or access to its internal parameters. Especially, we design a comprehensive evaluation pipeline tailored to this critical yet overlooked problem. Through extensive experiments on four publicly available benchmarking datasets and five representative survival models, we demonstrate the superiority of *SurvUnc* across multiple evaluation scenarios, including selective prediction, misprediction detection, and out-of-domain detection. Our results highlight the effectiveness of *SurvUnc* in enhancing model interpretability and reliability, paving the way for more trustworthy survival predictions in real-world applications.

CCS Concepts

• **Mathematics of computing** → **Survival analysis**; • **Applied computing** → **Health informatics**; • **Computing methodologies** → **Uncertainty quantification**; **Artificial intelligence**.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
KDD '25, August 3–7, 2025, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1454-2/2025/08
<https://doi.org/10.1145/3711896.3737140>

Keywords

Survival analysis, uncertainty quantification, meta model, out-of-domain detection

ACM Reference Format:

Yu Liu, Wei Yao Tao, Tong Xia, Simon Knight, and Tingting Zhu. 2025. *SurvUnc: A Meta-Model Based Uncertainty Quantification Framework for Survival Analysis*. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3711896.3737140>

1 Introduction

Survival analysis, or equivalently time-to-event analysis, which aims to estimate when an event of interest is likely to occur, has received substantial attention from various fields, such as predicting patient death or disease risk in healthcare, product lifespan in manufacturing, and customer churn in finance [49]. Since the event of interest is not always observed (e.g., many patients are lost to follow-up), survival data are frequently right-censored (i.e., the event such as death occurs after the censoring) [8], making survival analysis a more complex problem compared to traditional regression and classification tasks.

Driven by the advancements of machine learning and deep learning in recent years, several models have been proposed for survival analysis [52]. For example, RSF [22] adapts random forest for this purpose. Built upon the linear Cox proportional hazard (CoxPH) model [6], DeepSurv [25] and SurvTRACE [50] apply fully connected networks and transformers to model the nonlinear relationship between covariates/features and the hazard rate of an event. DeepHit [30] and DSM [39] take a different approach by directly modeling the event occurrence probability using neural networks.

Despite their advancements, these survival models are subject to the inherent noise in the data and the lack of knowledge on the optimal modeling approach, making them inapplicable for unseen data in practice [11]. The lack of uncertainty quantification for survival models significantly reduces their reliability in risk-sensitive applications such as healthcare [35, 36, 53]. While much progress has been made in quantitatively measuring the reliability of a model in prediction [47], most uncertainty quantification studies focus on classification and regression tasks. Bayesian approaches, such as Bayesian Neural Networks (BNNs) and Monte Carlo Dropout (MC-Dropout) [14], have been explored for uncertainty estimation

in these domains [36, 43]. Additionally, meta-models have been developed to quantify classification uncertainty in a supervised setting [44]. However, uncertainty quantification for survival models remains largely unexplored, presenting critical challenges that need to be addressed:

- **Survival Model-Agnostic Integration.** The diversity of survival models requires a robust standalone uncertainty quantification framework, where the latter is capable of integrating into any model without modifications. For example, uncertainty quantification methods with BNNs and MC-Dropout are incompatible with survival models such as RSF [22]. Thus, model-agnostic integration is essential to accommodate the wide range of existing and emerging survival models.
- **Absence of Ground-Truth Survival Curves.** Survival curves describe the survival probability of an event (e.g., death) not occurring by a specific time. However, in real-world settings, the “true” survival curves are unknown, and for censored samples, the actual event times remain uncertain. This absence of ground truth makes it infeasible to directly apply supervised uncertainty quantification methods to survival models, posing a significant challenge in reliability assessment.
- **Lack of Standardized Evaluation Protocols.** Despite the increasing focus on uncertainty quantification, there is no universally accepted evaluation framework for assessing the quality of uncertainty estimates in survival models. Beyond the methodological challenges in developing uncertainty quantification methods, establishing robust and systematic evaluation protocols remains an open problem.

To address the aforementioned challenges, we propose *SurvUnc*, a post-hoc meta-model based uncertainty quantification framework for survival models. More specifically, we develop a lightweight meta-model that acts as an “observer” on top of any existing survival model (base model). Sharing the same input covariates as the survival model, the meta-model is trained to estimate the uncertainty of the prediction of the survival model, which requires neither modifications to the survival model nor access to its architecture and parameters. Especially, motivated by the classic concordance idea in survival analysis [17], we design an anchor-based learning strategy to address the absence of ground-truth survival curves in meta-model optimization. Our strategy (i) selects a group of uncensored samples as anchors; (ii) evaluates whether the survival model correctly ranks the survival probabilities of samples relative to these anchors; and (iii) incorporates the evaluation outcome into the meta-model training to quantify the uncertainty of the survival model’s predictions. Furthermore, we refine and adapt existing evaluation protocols from uncertainty quantification studies in classification and regression to suit the survival analysis context, ensuring a more standardized assessment of uncertainty estimation in survival models.

In summary, our key contributions are as follows:

- We propose *SurvUnc*, a post-hoc meta-model based uncertainty quantification framework for survival models. Notably, it requires no modifications or access to the model’s parameters. To the best of our knowledge, *SurvUnc* is the first model-agnostic framework capable of quantifying the uncertainty of predictions in any survival model.
- We develop an anchor-based learning strategy that leverages the concordance concept of survival analysis to construct the meta-model training set for optimization, thereby providing deeper insights into understanding the uncertainty in survival analysis.
- We design systematic evaluation protocols to assess the quality of quantified uncertainty in survival models, including selective prediction, misprediction detection, and out-of-domain (OOD) detection tasks. These protocols offer an evaluation reference for future research in uncertainty estimation for survival analysis.
- We validate *SurvUnc* on multiple survival models and datasets. Our extensive experiments demonstrate that *SurvUnc* significantly outperforms baseline methods across various tasks, underscoring its effectiveness and robustness in uncertainty quantification for survival models.

2 Related Work

2.1 Survival Analysis

Survival analysis is concerned with modeling event occurrence in the presence of censoring, and it primarily involves two interrelated quantities: (i) *hazard function*, which represents the instantaneous rate of event occurrence at a given time point, conditional on the event not having occurred prior to that time; and (ii) *survival function*, (a.k.a. the *survival curve*), which denotes the probability that the event has not occurred by a specific time. Survival probability is determined by the cumulative hazard up to that time. Survival models typically focus on estimating either the hazard function or the survival function [52].

Traditionally, statistical methods have been widely explored for survival analysis. Kaplan-Meier estimator [24] firstly defines the survival function based on empirical estimation of the survival data. CoxPH model [6] assumes that the log of the hazard is a linear combination of covariates, and the ratio between the hazards of two samples is constant. Furthermore, several machine learning methods have been adapted for survival analysis [49]. One of the most representative examples is RSF [22], which extends the random forest to survival analysis.

Owing to the ability to capture high-level non-linear interactions, neural network-based deep learning models have gained popularity in survival analysis [52]. Early models, such as the Farggi-Simon model [12] and DeepSurv [25], replace the linear combination in the Cox model with fully connected networks. Furthermore, recurrent neural networks and Transformers are also explored for modeling hazard functions in DRSA [42] and SurvTRACE [50], respectively. On the other hand, DeepHit [30] and DSM [39] directly target the survival function using fully connected networks and parametric distributions. A recent work, MOTOR [45] introduces a Transformer-based foundation model pretrained on time-to-event data for survival analysis. However, the uncertainty associated with predictions from such survival models remains unquantified.

In addition, several Bayesian methods have been developed for survival analysis. For instance, the Gaussian process is firstly incorporated into the hazard function for a semi-parametric Bayesian model in [13], while DSA [38, 41] employs a deep latent variable model within a Bayesian framework to analyze survival data. BNNs [31, 32, 35] and deep Gaussian processes [1] further integrate neural

network architectures with Bayesian techniques for survival prediction. Although Bayesian methods inherently provide uncertainty quantification through iterative sampling or posterior inference, they are typically confined to probabilistic models and often struggle with generalization or integration with other survival models. This limitation can reduce their practical reliability in real-world applications. Overall, accurately quantifying prediction uncertainty in survival models remains a significant challenge.

2.2 Uncertainty Quantification

Uncertainty quantification aims to quantitatively measure the reliability and confidence of a model's predictions, and can be categorized into *intrinsic* and *post-hoc (extrinsic)*, depending on whether uncertainty estimation is integrated within model training [44].

Intrinsic uncertainty quantification methods integrate uncertainty estimation directly into the model learning process, often leveraging Bayesian approaches to naturally generate uncertainty estimates as part of their predictions [15]. For instance, various studies employ BNNs with techniques such as variational inference and Laplace approximation to quantify uncertainty in both classification and regression tasks [34, 40]. In addition, evidential deep learning methods, which parameterize a Dirichlet distribution on the neural network outputs, have been used to quantify uncertainty in classification problems [37]. However, these methods are often computationally intensive and based on strong underlying assumptions. Furthermore, intrinsic methods cannot directly quantify the uncertainty in survival models.

In comparison, post-hoc uncertainty quantification methods separate the uncertainty estimation from the model prediction, without affecting the model learning process. Prominent examples include MC-Dropout¹ [14] and Deep Ensembles [29], widely used in various domains [18]. MC-Dropout estimates uncertainty by calculating the variance of multiple forward predictions with dropout layers activated during inference. Deep Ensembles define uncertainty as the prediction variance across multiple independent models trained from different random initializations, while Hyper-batch Ensembles [51] further consider different hyperparameters in the ensemble. Another promising paradigm is to build auxiliary or meta-models to quantify the uncertainty of the base model with respect to its original task, as studied in classification [4, 10, 44] and regression [54]. However, the effectiveness of these methods in survival analysis remains unclear due to the unique challenges posed by censored data and the absence of ground-truth survival curves, which are fundamental to meta-model methods. Our work addresses this gap by introducing a novel post-hoc uncertainty quantification method specifically tailored for survival models.

3 Proposed Framework

3.1 Preliminaries & Problem Definition

3.1.1 Survival Analysis. Firstly, we define a survival dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i, \delta_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the set of covariates/features of an individual; $t_i \in \mathbb{R}^+$ is the time to event or censoring as indicated by the indicator $\delta_i = 1$ or $\delta_i = 0$, respectively. N refers to the number of samples and d is the dimension of covariates.

¹MC-Dropout requires a dropout layer to be designed in the quantified model.

Moreover, the probability that an individual with covariates \mathbf{x} will experience the event at time t , i.e., the probability density function of survival times, is denoted as $p(t|\mathbf{x}) = \mathbb{P}(T = t|X = \mathbf{x})$. Consequently, the survival function is represented as:

$$S(t|\mathbf{x}) = \mathbb{P}(T > t|X = \mathbf{x}) = 1 - \int_0^t p(z|\mathbf{x})dz, \quad (1)$$

which signifies the probability that the event does not occur until time t . The hazard function is defined as:

$$h(t|\mathbf{x}) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t, X = \mathbf{x})}{\Delta t} = \frac{p(t|\mathbf{x})}{S(t|\mathbf{x})}, \quad (2)$$

which indicates the probability that the event will occur at time t , given that the event has not occurred before. Consequently, the survival model $F(\cdot)$ is developed to learn $S(t|\mathbf{x})$ and $h(t|\mathbf{x})$ with covariates \mathbf{x} as input, based on the survival dataset \mathcal{D} .

3.1.2 Uncertainty Quantification. In the context of machine learning and deep learning, the total uncertainty encompasses two types of uncertainty: (i) reducible *epistemic uncertainty* (a.k.a. model uncertainty) caused by the model's limited knowledge due to insufficient training data, e.g., unseen data samples; and (ii) irreducible *aleatoric uncertainty* (a.k.a. data uncertainty) caused by the inherent noise and stochastic nature in data [20]. Based on the preliminaries, we formally define the problem of uncertainty quantification in survival analysis as follows.

PROBLEM 1. Uncertainty Quantification in Survival Analysis. Given a survival model $F(\cdot)$ trained on a survival dataset \mathcal{D} , the uncertainty quantification problem is to learn a function $U(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$, such that based on $F(\cdot)$ and \mathcal{D} , it produces an uncertainty score that quantifies the predictive uncertainty of $F(\cdot)$ for a new coming sample.

3.2 Framework Overview

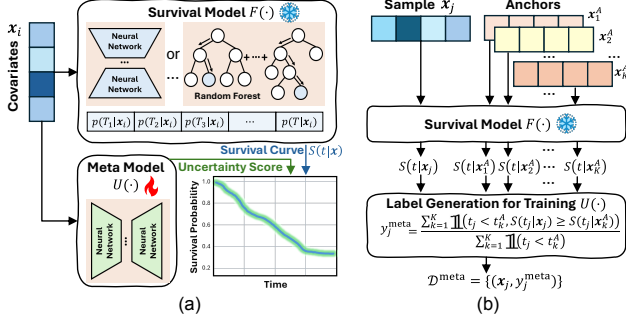
Figure 1 illustrates an overview of our proposed framework, SurvUnc, for the uncertainty quantification problem in survival analysis. Specifically, we propose a meta-model based framework to quantify the uncertainty of survival models in a post-hoc manner. Given a pretrained survival model $F(\cdot)$ (e.g., neural network-based model, or random forest-based model), we further develop a meta-model $U(\cdot)$ to quantify its uncertainty. Notably, the meta-model shares the same covariate input as the survival model, and only the parameters of the meta-model are optimized in the learning process. In general, the architecture of the meta-model can be arbitrarily chosen; in our experiments, we explore two different variants of $U(\cdot)$, as shown in later experiments. To overcome the absence of ground truth of survival curves for meta-model training, we design an anchor-based learning strategy to construct a meta-model training set, $\mathcal{D}^{\text{meta}}$, derived from the survival model training set \mathcal{D} and the pretrained survival model $F(\cdot)$, as illustrated in Figure 1(b).

3.3 Meta-Model Learning

To construct a labeled meta-model training set, we leverage the concordance concept from survival analysis in the medical domain [17] and pairwise ranking idea in information retrieval domain [33]. The core idea in medical language is that a patient living shorter should have a lower survival probability at any given time

Table 1: Dataset statistics. “real” and “categorical” denote the number of real-valued and categorical covariates in the datasets, respectively. Minimum, maximum, and mean values for event durations and censoring times are reported.

Dataset	#Instances	#Events (%)	#Censored (%)	#Covariates (real, categorical)	Event Duration			Censoring Time		
					min	max	mean	min	max	mean
FLCHAIN	6,524	1,962 (30.1%)	4,562 (69.9%)	7 (3, 4)	0	4998	2137.9	1	5166	4296.7
SUPPORT	8,873	6,036 (68.0%)	2,837 (32.0%)	14 (8, 6)	3	1944	205.4	344	2029	1059.8
SEER-BC	323,772	85,258 (26.3%)	238,514 (73.7%)	18 (4, 14)	1	121	41.7	1	121	85.4
SAC3	100,000	62,798 (62.8%)	37,202 (37.2%)	49 (49, 0)	0.1	100	24.6	0.1	100	45.5

**Figure 1: Illustration of our proposed framework SurvUnc.** (a) The pipeline of post-hoc meta-model based uncertainty quantification for survival models, and (b) the anchor-based learning strategy for meta-model optimization.

compared to a patient living longer. Therefore, the uncertainty in a survival model’s predictions can be quantified by evaluating its ability to correctly rank the survival probabilities of samples relative to each other.

Intuitively, comparing a sample against some referenced points enhances the quality of the quantified uncertainty. For efficiency consideration, we design an anchor-based learning strategy. Following the literature we assume that censoring occurs completely at random [30, 39]. Specifically, we randomly select K uncensored samples from the survival model training set \mathcal{D} as anchors, denoted by $\mathcal{D}^A = \{(x_k^A, t_k^A, \delta_k^A = 1)\}_{k=1}^K$. For the j -th uncensored sample $(x_j, t_j, \delta_j = 1) \in \mathcal{D}$, we treat it as a training instance for the meta-model, and calculate its uncertainty label as follows:

$$y_j^{\text{meta}} = \frac{\sum_{k=1}^K \mathbb{1}(t_j < t_k^A, S(t_j|x_j) \geq S(t_j|x_k^A))}{\sum_{k=1}^K \mathbb{1}(t_j < t_k^A)}, \quad (3)$$

which measures the proportion of incorrectly ordered pairs by the survival model $F(\cdot)$, resulting in an uncertainty score ranging from 0 to 1 for the predicted sample. A higher uncertainty score indicates a less reliable prediction. Here $\mathbb{1}(\cdot)$ is an indicator function that returns 1 if the specified conditions are met and 0 otherwise. $S(t_j|x_j)$ and $S(t_j|x_k^A)$ denote the survival probabilities estimated by $F(\cdot)$ for samples with covariates x_j and x_k^A at time t_j , respectively. As a result, we construct the meta-model training set $\mathcal{D}^{\text{meta}} = \{(x_j, y_j^{\text{meta}})\}_{j=1}^{N'}$ with N' labeled samples.

Algorithm 1: The learning procedure of SurvUnc

Input: Survival data $\mathcal{D} = \{(x_i, t_i, \delta_i)\}_{i=1}^N$;
Pretrained survival model $F(\cdot)$; Anchor number K

Output: Uncertainty quantification model $U(\cdot)$

- 1 Randomly sample K anchors from \mathcal{D} , i.e., \mathcal{D}^A
 - 2 Initialize the model $U(\cdot)$
 - 3 $\mathcal{D}^{\text{meta}} \leftarrow \emptyset$
 - 4 **for** $(x_j, t_j, \delta_j = 1) \in \mathcal{D}$ **do**
 - 5 Obtain label y_j^{meta} based on (3) and \mathcal{D}^A
 - 6 $\mathcal{D}^{\text{meta}} \leftarrow \mathcal{D}^{\text{meta}} \cup \{(x_j, y_j^{\text{meta}})\}$
 - 7 Train the model $U(\cdot)$ with $\mathcal{D}^{\text{meta}}$
- Usage:** $u_i = U(x_i)$
-

Based on the constructed meta-model training set $\mathcal{D}^{\text{meta}}$, we explore two different architectures of the meta-model $U(\cdot)$ with sample covariates as input, including a multi-layer perceptron (MLP) model and a random forest model, referred to as **SurvUnc-MLP** and **SurvUnc-RF**, respectively. After training on $\mathcal{D}^{\text{meta}}$, the trained SurvUnc models provide uncertainty scores for the predictions from the corresponding survival model. Algorithm 1 summarizes the learning procedure of SurvUnc. Note that during inference, the event or censoring time for a test sample is unknown, requiring the meta-model $U(\cdot)$ to estimate predictive uncertainty solely from the covariates.

4 Experiments

In this section, we show that our proposed SurvUnc framework effectively quantifies uncertainty in survival models and is applicable in different scenarios. Specifically, we conduct experiments to answer the following questions:

- **RQ1:** How does SurvUnc perform w.r.t. total uncertainty quantification, particularly in distinguishing between correctly and incorrectly predicted samples, compared to existing methods?
- **RQ2:** How does SurvUnc perform w.r.t. epistemic uncertainty quantification, particularly in detecting OOD data, when compared to existing methods?
- **RQ3:** How do different hyperparameter settings, such as meta-model structure and the number of anchors, affect the performance of SurvUnc?

4.1 Experimental Settings

4.1.1 Datasets. We experiment with three real-world and one synthetic, publicly available survival analysis datasets:

- **FLCHAIN** is a study of the relationship between free light chains and mortality in an elderly population [9].
- **SUPPORT** contains a dataset that aims to improve the care for seriously ill patients by understanding their prognosis and treatment preferences [5].
- **SEER-BC** is derived from the SEER database [21], which contains survival information for oncology patients in the U.S. Following the processing steps in [50], we select breast cancer patients to create the SEER-BC dataset. For the later OOD detection experiments, we curate a SEER-HD dataset, comprising the same number of patients as SEER-BC but diagnosed with heart disease.
- **SAC3** is a synthetic dataset from [26], which simulates survival times by sampling from a combination of three hazard functions. Thus, this dataset provides the ground-truth survival curves.

Table 1 presents the basic statistics of all datasets. For the SUPPORT dataset, we retain the original training/test split and reserve 20% of the training set for validation. For all other datasets, we randomly split into training, validation, and test sets in a ratio of 6:2:2.

4.1.2 Survival Models. To evaluate the effectiveness as well as robustness of our proposed SurvUnc framework on survival models, we select five representative survival models for uncertainty quantification, including three deep learning models of **DeepSurv** [25], **DeepHit** [30], **DSM** [39], one machine learning model of **RSF** [22], and one Bayesian-based model of **BNNSurv** [31]. For each survival model, we optimize hyperparameters so that the reported performance on datasets is comparable to those published in the literature, thereby validating the correctness of pretrained survival models.

4.1.3 Uncertainty Quantification Baselines. Given the lack of prior studies on model-agnostic uncertainty quantification for survival models, we adapt two widely used uncertainty quantification methods of **MC-Dropout** [14] and **Deep Ensembles (Ensemble)** [29] to survival analysis. For MC-Dropout, we set the number of forward passes to 100. For Ensemble, we train 10 survival models with different random initializations. Note that the RSF model does not support the dropout mechanism, making MC-Dropout incompatible with it. Unlike MC-Dropout and Ensemble, which are general uncertainty quantification methods applicable across models, BNNSurv is a specialized Bayesian survival model that inherently estimates uncertainty but cannot be integrated with other survival models for uncertainty quantification.

4.1.4 Tasks & Metrics. As discussed before, standardized evaluation protocols for uncertainty quantification methods with survival models remain lacking. To address this, we extend existing evaluation methodologies from uncertainty quantification studies to the survival analysis domain, incorporating multiple tasks and metrics.

Selective prediction evaluates how well an uncertainty quantification method differentiates between correct and incorrect predictions. A good uncertainty quantification method should improve survival model performance by selectively discarding samples with high uncertainty. To evaluate this, we use two widely adopted survival analysis metrics.

- **Time Dependent Concordance Index (C^{td}).** This metric measures the model discrimination power by comparing the relative survival probabilities across all pairs of samples in the test set

[2], defined as

$$C^{td} = \frac{\sum_{i=1}^n \sum_{j=1}^n \delta_i \cdot \mathbb{1}(t_i < t_j, S(t_i|\mathbf{x}_i) < S(t_i|\mathbf{x}_j))}{\sum_{i=1}^n \sum_{j=1}^n \delta_i \cdot \mathbb{1}(t_i < t_j)}.$$

The range of C^{td} is $[0, 1]$, and a larger value indicates better model prediction. $C^{td}=0.5$ corresponds to a random prediction.

- **Integral Brier Score (IBS).** This metric is an integral of the Brier score at all time points, while the Brier score calculates the mean squared error between predicted survival probability and binary observation at a given time [27], defined as

$$BS(t) = \frac{1}{n} \sum_{i=1}^n \left[\frac{(0 - S(t|\mathbf{x}_i))^2 \cdot \mathbb{1}(t \geq t_i, \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1 - S(t|\mathbf{x}_i))^2 \cdot \mathbb{1}(t < t_i)}{\hat{G}(t)} \right],$$

where $\hat{G}(t)$ is the censoring survival function using Kaplan-Meier estimator [24]. Unlike C^{td} , which relies on pairwise comparisons, IBS assesses each sample independently and provides an overall measure of model accuracy, where lower values indicate better performance. The range of IBS is $[0, 1]$.

Misprediction detection examines whether uncertainty estimates align with actual prediction errors. To assess this, we use the following metric:

- **Pearson Correlation Coefficient ρ .** This metric measures the linear correlation between the quantified uncertainty scores and IBSs of test samples, i.e., to evaluate whether the uncertainty quantification method can distinguish correctly and incorrectly predicted samples by survival models. The range of ρ is $[-1, 1]$, while we expect a positive correlation ($\rho \geq 0$) here, i.e., samples with larger uncertainty have larger IBS.

OOD detection evaluates the ability of uncertainty quantification methods to distinguish in-distribution (IND) from OOD samples. We use the following two metrics:

- **AUROC and AUPRC.** We adopt the area under the receiver operating curve (AUROC) and the area under the precision-recall curve (AUPRC) to evaluate the performance in OOD detection experiments. Specifically, IND test samples are labeled as the negative class, while OOD samples are labeled as the positive class [44]. Both metrics are in the range of $[0, 1]$, and a higher value indicates better performance.

4.1.5 Implementation Details. For the SurvUnc-RF method, the hyperparameters `n_estimators`, `min_samples_leaf`, `min_samples_split` are uniformly set to 100, 5 and 10, respectively, across all survival models and datasets. Similarly, for the SurvUnc-MLP method, the learning rate and the hidden layers are simply set to 0.001 and [32, 32], respectively. The number of anchors is usually set to 50 for robust performance. All methods undergo 100 bootstrap resampling iterations on the test set, with both the mean value and standard deviation reported. We also conduct the Wilcoxon signed-rank test to confirm the statistical significance of our results. All experiments were run on an RTX 6000 GPU with 32GB RAM. The implementation is done in PyTorch, and training the SurvUnc framework with a pretrained survival model on the largest SEER-BC dataset takes less than 5 minutes, making it efficient enough for deployment. The

Table 2: C^{td} of survival models under different discarding percentages (10%, 30%, 50%), determined by different uncertainty quantification (UQ) methods across datasets. The best results are in bold, and the last row of each group shows relative improvement compared with the best baseline by 100 runs of experiments. p -value < 0.001 is true for all results without *.

Survival Models	UQ Methods	FLCHAIN			SUPPORT			SEER-BC			SAC3		
		10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
DeepSurv	MC-Dropout	0.782	0.745	0.685	0.600	0.576	0.563	0.857	0.846	0.832	0.772	0.778	0.793
	Ensemble	0.786	0.755	0.710	0.607	0.598	0.576	0.863	0.857	0.842	0.783	0.795	0.807
	SurvUnc-RF	0.856	0.907	0.941	0.635	0.690	0.757	0.904	0.938	0.961	0.792	0.822	0.855
	SurvUnc-MLP	0.839	0.894	0.935	0.637	0.695	0.762	0.904	0.938	0.961	0.797	0.830	0.862
	Improv.	8.9%	20.1%	32.5%	4.9%	16.2%	32.3%	4.8%	9.5%	14.1%	1.8%	4.4%	6.8%
DeepHit	MC-Dropout	0.784	0.762	0.733	0.648	0.657	0.656	0.908	0.937	0.953	0.797	0.811	0.836
	Ensemble	0.855	0.902	0.934	0.637	0.637	0.637	0.896	0.923	0.944	0.805	0.822	0.842
	SurvUnc-RF	0.855*	0.908	0.940	0.639	0.664*	0.693	0.913	0.945	0.961	0.811	0.838	0.866
	SurvUnc-MLP	0.855*	0.909	0.940	0.639	0.660*	0.692	0.914	0.945	0.961	0.817	0.845	0.876
	Improv.	0.0%	0.8%	0.6%	-1.4%	1.1%	5.6%	0.7%	0.9%	0.8%	1.5%	2.8%	4.0%
DSM	MC-Dropout	0.780	0.741	0.687	0.613	0.614	0.613	0.866	0.854	0.832	0.793	0.776	0.770
	Ensemble	0.787	0.755	0.715	0.616	0.605	0.581	0.872	0.868	0.854	0.802	0.813	0.828
	SurvUnc-RF	0.854	0.907	0.941	0.640	0.685	0.725	0.910	0.943	0.960	0.813	0.837	0.862
	SurvUnc-MLP	0.852	0.906	0.940	0.641	0.685	0.731	0.910	0.943	0.959	0.818	0.844	0.871
	Improv.	8.5%	20.1%	31.6%	4.1%	11.6%	19.2%	4.4%	8.6%	12.4%	2.0%	3.8%	5.2%
RSF	MC-Dropout	-	-	-	-	-	-	-	-	-	-	-	-
	Ensemble	0.790	0.777	0.745	0.648	0.662	0.684	0.878	0.874	0.863	0.649	0.663	0.677
	SurvUnc-RF	0.840	0.897	0.933	0.663	0.708	0.750	0.908	0.941	0.959	0.658	0.692	0.738
	SurvUnc-MLP	0.820	0.854	0.892	0.656	0.689	0.721	0.913	0.943	0.960	0.653	0.685	0.731
	Improv.	6.3%	15.4%	25.2%	2.3%	6.9%	9.6%	4.0%	7.9%	11.2%	1.4%	4.4%	9.0%
BNNSurv	Bayesian	0.773	0.732	0.670	0.623	0.648	0.687	0.847	0.836	0.805	0.719	0.740	0.771
	SurvUnc-RF	0.848	0.902	0.936	0.650	0.701	0.764	0.891	0.932	0.955	0.727	0.765	0.808
	SurvUnc-MLP	0.843	0.899	0.934	0.648	0.699	0.760	0.893	0.932	0.955	0.727	0.767	0.815
	Improv.	9.7%	23.2%	39.7%	4.3%	8.2%	11.2%	5.4%	11.5%	18.6%	1.1%	3.6%	5.7%

implementation code and dataset are available at the given link². Details on the experimental settings can be found in Appendix A.

4.2 Selective Prediction (RQ1)

To investigate the effectiveness of total uncertainty quantification, we first conduct the selective prediction experiments [48, 55]. Specifically, we discard a portion of test samples by their uncertainty scores and then compute evaluation metrics on the remaining data. To better reflect real-world scenarios where testing samples have not yet been censored, we discard only uncensored samples, and subsequent analyses also focus on uncensored ones. It is expected that the performance will improve as more high-uncertainty samples are discarded. This evaluation is essential in realistic scenarios, where low-uncertainty predictions are retained, while high-uncertainty (less reliable) predictions are flagged for expert inspection.

Table 2 presents the selective prediction results for C^{td} across five survival models, evaluated at different discarding percentages (10%, 30%, 50%). Since BNNSurv inherently provides uncertainty estimation, we exclude MC-Dropout and Ensemble implementations for this model. Across all survival models, datasets and discarding

percentages, our proposed SurvUnc framework generally outperforms the baselines with an average improvement of over 8%, which demonstrates the effectiveness of the meta-model based uncertainty quantification framework and the anchor-based learning strategy. Especially, our approach aligns with the core objective of survival analysis, i.e., maintaining correct risk ordering, whose uncertainty scores better reflect this clinical need by focusing on ranking reliability. In comparison, traditional methods focus on predictive variability, which do not account for the time-dependent nature of survival predictions. This leads to two main issues including that they may flag samples with high variance but correct rankings as “uncertain” and they often miss samples with small variance but incorrect relative rankings. Notably, the relatively weak performance of BNNSurv’s inherent uncertainty estimation suggests that Bayesian methods may struggle to capture meaningful uncertainty scores in survival analysis. The marginal improvement observed with DeepHit could be explained by its use of ranking loss, which also leverages the concordance concept, thereby making MC-Dropout and Ensemble methods based on its output more effective. Additionally, SurvUnc-RF and SurvUnc-MLP demonstrate comparable performance across different survival models and datasets, further validating the robustness and generalizability of SurvUnc.

²<https://github.com/liuyuaa/SurvUnc>

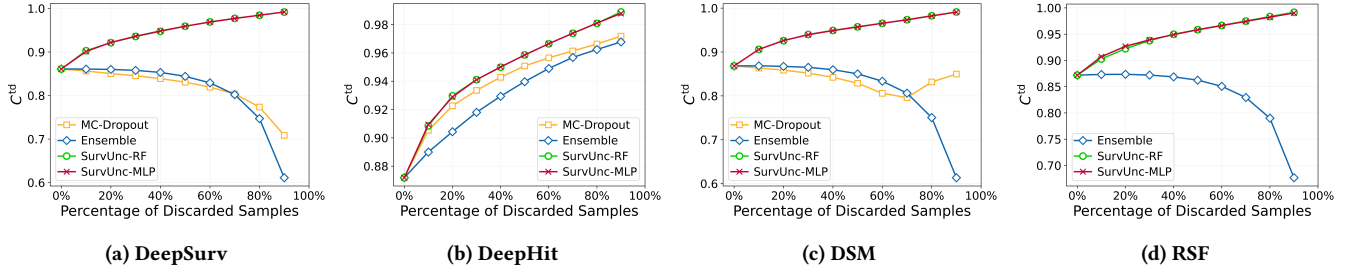


Figure 2: C^{td} of four survival models of (a) DeepSurv, (b) DeepHit, (c) DSM and (d) RSF on SEER-BC dataset with different percentages of samples discarded according to uncertainty scores from different UQ methods. A consistent upward trend is expected as the percentage of discarded samples increases. Error bars are omitted for better visualization.

Table 3: Misprediction detection results, i.e., Pearson correlation coefficient between uncertainty scores and IBSs of samples.

Datasets	FLCHAIN				SUPPORT				SEER-BC				SAC3			
UQ Methods	DeepSurv	DeepHit	DSM	RSF	DeepSurv	DeepHit	DSM	RSF	DeepSurv	DeepHit	DSM	RSF	DeepSurv	DeepHit	DSM	RSF
MC-Dropout	-0.674	-0.042	-0.579	-	-0.299	0.431	-0.123	-	-0.293	0.442	-0.644	-	-0.111	-0.157	-0.329	-
Ensemble	-0.361	0.099	-0.467	-0.258	-0.109	0.511	-0.277	0.203	-0.168	0.391	-0.209	-0.264	-0.036	-0.017	-0.085	0.129
SurvUnc-RF	0.688	0.084	0.718	0.687	0.657	0.590	0.516	0.544	0.683	0.331	0.698	0.607	0.426	0.315	0.342	0.436
SurvUnc-MLP	0.590	0.053	0.700	0.352	0.652	0.509	0.468	0.405	0.671	0.323	0.677	0.666	0.638	0.510	0.554	0.391

Figure 2 shows the performance comparison with fine-grained discarding percentages on the largest dataset, SEER-BC. We observe that the performance of both SurvUnc-RF and SurvUnc-MLP improves as the discarding percentage increases, indicating that predictions become more reliable as high-uncertainty samples are excluded, consistent with the expectation of selective prediction experiments. However, for baselines quantifying the uncertainty of DeepSurv, DSM and RSF (see Figure 2(a), (c) and (d)), the results are opposite, suggesting that these baselines fail to provide meaningful uncertainty quantification for these survival models. We also conduct the experiments using the IBS metric, which yields similar conclusions for most results (see Appendix B).

4.3 Misprediction Detection (RQ1)

We conduct the misprediction detection experiments to identify the correctly and incorrectly predicted samples using quantified uncertainty scores [44], where the mispredicted samples are viewed as in-distribution hard samples for survival models. Table 3 summarizes the Pearson correlation coefficients between the IBSs and quantified uncertainty scores from different uncertainty quantification methods for each survival model across four datasets. Results with BNNSurve are provided in Appendix B. As observed, both SurvUnc-RF and SurvUnc-MLP achieve significantly higher and consistently positive correlations across all survival models and datasets, demonstrating their reliability in uncertainty quantification. In comparison, the performance of baselines is inconsistent, and several negative or weak correlation results are notable. The results indicate that traditional uncertainty quantification methods like MC-Dropout and Ensemble struggle to extend effectively to survival analysis, further demonstrating the effectiveness and importance of SurvUnc for survival analysis.

To further analyze the effectiveness of total uncertainty quantification, Figure 3 presents a comparison between predicted uncertainty scores and IBSs for the DeepSurv model using SurvUnc-RF

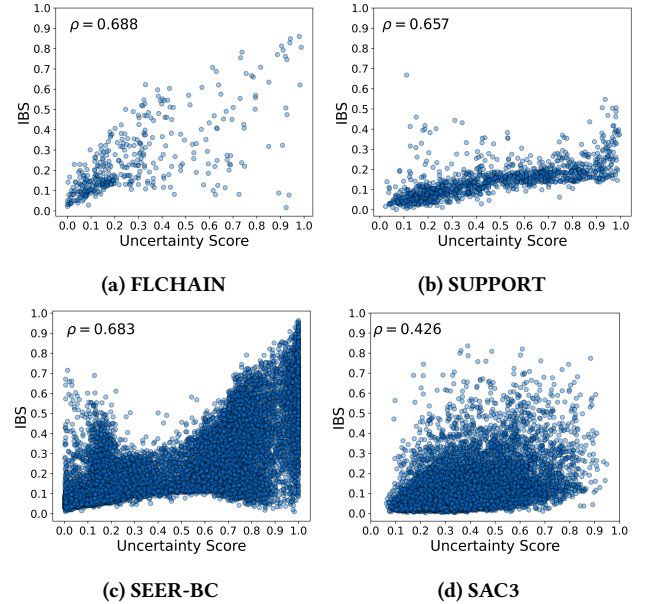


Figure 3: Predicted uncertainty scores versus IBSs from DeepSurv quantified by SurvUnc-RF across samples on (a) FLCHAIN, (b) SUPPORT, (c) SEER-BC and (d) SAC3 datasets.

on four datasets. The scatter plots show a clear positive correlation between quantified uncertainty and prediction performance. Notably, despite the different prediction distributions of IBSs across the four datasets, Our SurvUnc framework can effectively differentiate between relatively hard and easy in-distribution samples by assigning accurate uncertainty scores for each dataset.

As the absence of ground-truth survival curves poses a significant challenge to uncertainty quantification for survival models,

Table 4: OOD detection comparison with AUROC and AUPRC on the SEER dataset. Since the Bayesian-based model BNNSurv inherently provides uncertainty estimates and cannot be integrated with other survival models, the corresponding columns are left blank.

Models	DeepSurv		DeepHit		DSM		RSF		BNNSurv	
UQ Methods	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
MC-Dropout	0.358	0.398	0.524	0.527	0.445	0.455	-	-	-	-
Ensemble	0.493	0.481	0.499	0.516	0.558	0.559	0.516	0.514	-	-
Bayesian	-	-	-	-	-	-	-	-	0.483	0.500
SurvUnc-RF	0.621	0.581	0.631	0.574	0.638	0.581	0.625	0.577	0.627	0.627
SurvUnc-MLP	0.638	0.604	0.657	0.599	0.667	0.624	0.643	0.577	0.636	0.634

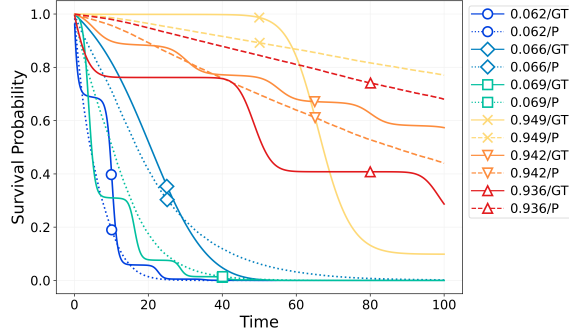


Figure 4: Survival curve comparison of high-uncertainty and low-uncertainty samples on SAC3 dataset, quantified by SurvUnc-RF on DeepSurv. “GT” (with solid lines) and “P” (with dashed lines) denote “Ground Truth” and “Predicted”, respectively, and the values in legend are uncertainty scores.

we introduce the synthetic SAC3 dataset with ground-truth survival curves available for validation. Given a pretrained DeepSurv model on the SAC3 dataset, we select the three samples with the highest uncertainty and the three with the lowest uncertainty, as ordered by SurvUnc-RF, for visualization. As shown in Figure 4, the predicted survival curves (in dashed lines with cold colors) for low-uncertainty samples closely match the ground-truth survival curves (in solid lines with cold colors). In contrast, for high-uncertainty samples, there is a marked discrepancy between the predicted and ground-truth survival curves. Our expanded analysis confirms that samples with high uncertainty predictions generally exhibit longer survival times compared to those with low uncertainty. The reason is current survival models tend to accumulate prediction errors when estimating long-term survival patterns, resulting in higher uncertainty estimates. These results further highlight the effectiveness of SurvUnc in quantifying total uncertainty associated with survival curves.

4.4 OOD Detection (RQ2)

In addition to the quantification of total uncertainty, the quantification of epistemic uncertainty is emphasized in reliable deployment to detect OOD data [44, 54]. Here, we conduct several OOD detection experiments to validate SurvUnc. As mentioned, we develop the SEER-HD dataset as the OOD counterpart to the SEER-BC

dataset. The SEER-HD dataset includes the same number of uncensored patients as the SEER-BC dataset but comprises patients with a different condition, specifically heart disease. For age at diagnosis, SEER-BC patients’ mean age is 60.52 (SD: 15.13) compared to SEER-HD patients with 74.84 (SD: 10.40). For time-to-event distribution, SEER-BC patients’ mean duration is 41.67 (SD: 29.91), compared to SEER-HD patients with 55.44 (SD: 33.34). These differences are statistically significant ($p < 0.001$) under Wilcoxon rank-sum test.

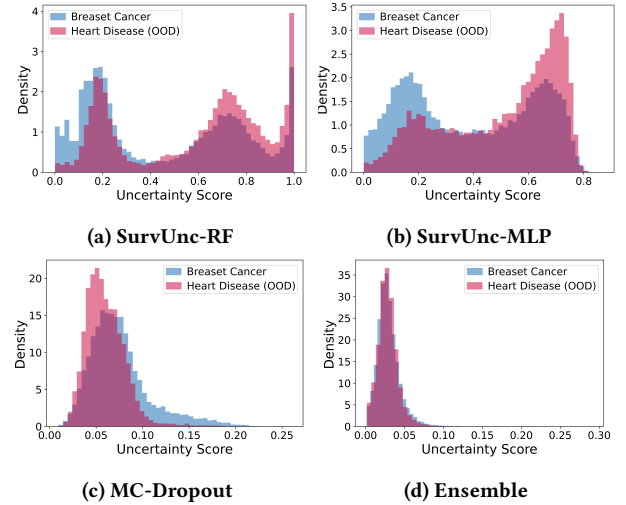


Figure 5: Uncertainty score distribution comparison of DeepSurv between IND (BC) and OOD (HD) samples from SEER dataset, with uncertainty scores from (a) SurvUnc-RF, (b) SurvUnc-MLP, (c) MC-Dropout and (d) Ensemble.

Firstly, we investigate whether the OOD data can be identified from the uncertainty score distribution. Based on DeepSurv, Figure 5 compares the uncertainty score distributions of both datasets across different quantification methods. The findings suggest that both SurvUnc-RF (Figure 5(a)) and SurvUnc-MLP (Figure 5(b)) can generally sense the OOD and IND data differently. Notably, the uncertainty score distribution for OOD data (shown in pink) is right-shifted in comparison to that of IND data (shown in blue), indicating higher uncertainty for more OOD samples. However, Ensemble (Figure 5(d)) fails to distinguish between the two distributions, as they appear nearly identical. For MC-Dropout, IND data

is generally quantified with even higher uncertainty scores than OOD data, highlighting its limitations in OOD detection.

Furthermore, we provide a quantitative evaluation of the OOD detection task in Table 4. By using uncertainty scores as logits to classify IND and OOD samples, SurvUnc demonstrates a significant performance improvement over baseline methods across all survival models. In particular, SurvUnc-MLP shows notable enhancements, achieving an 19.5%–31.7% increase in AUROC and an 11.6%–26.8% improvement in AUPRC. This epistemic uncertainty quantification capability can be attributed to the anchor-based learning strategy, which successfully infuses in-domain knowledge into the meta-model learning process.

4.5 Hyperparameter Study (RQ3)

Our proposed SurvUnc framework consists of two important hyperparameters, i.e., the meta-model structure and the number of anchors. In terms of the meta-model structure, we explore the random forest and MLP in former experiments, where both meta-models achieve consistently better performance than baselines with robustness achieved.

We investigate the impact of varying the number of anchors on selective prediction performance (C^{td}), as shown in Figure 6. Additional results with SurvUnc-MLP, provided in Appendix B, exhibit similar trends. The anchor-based learning strategy suggests that increasing the number of anchors can enhance the robustness of label calculation for the meta-model training set construction, thereby improving performance. For each discarding percentage, the performance stabilizes when using 10 or more anchors³. This indicates that the proposed SurvUnc framework can achieve efficient performance with a relatively small number of anchors.

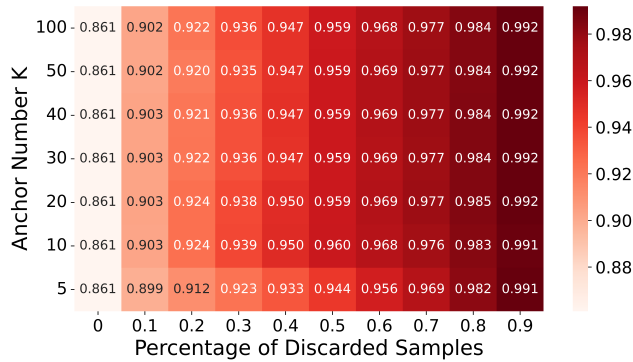


Figure 6: Varying anchor number K on selective prediction performance of SurvUnc-RF with DeepSurv on SEER-BC.

5 Conclusion

In this paper, we address the novel challenge of uncertainty quantification in survival analysis and propose SurvUnc, a meta-model based framework that incorporates an anchor-based learning strategy. We establish systematic evaluation protocols to assess uncertainty quantification methods on survival models. Extensive

³The slight variations are due to randomness in training.

experiments conducted across multiple datasets and survival models demonstrate the effectiveness and robustness of SurvUnc in uncertainty quantification. Furthermore, its model-agnostic design ensures compatibility with a wide range of survival models, offering valuable insights into uncertainty quantification for survival analysis. Future work will explore uncertainty quantification for survival analysis with competing events and time-varying covariates, as well as evaluate SurvUnc in the context of foundation model-based survival analysis [16, 23].

Acknowledgments

Some aspects of this work were funded through an award from the NIHR (AI award 02316). Tingting Zhu was supported by the Royal Academy of Engineering under the Research Fellowship scheme.

References

- [1] Ahmed M Alaa and Mihaela van der Schaar. 2017. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In *Advances in Neural Information Processing Systems*. 2326–2334.
- [2] L. Antolini, P. Boracchi, and E. Biganzoli. 2005. A Time-dependent Discrimination Index for Survival Data. *Statistics in Medicine* 24, 24 (2005), 3927–3944.
- [3] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive Interpretable Tabular Learning. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 35. 6679–6687.
- [4] Tongfei Chen, Jiri Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam. 2019. Confidence Scoring Using Whitebox Meta-models with Linear Classifier Probes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1467–1475.
- [5] Alfred F Connors, Neal V Dawson, Norman A Desbiens, William J Fulkerson, Lee Goldman, William A Knaus, Joanne Lynn, Robert K Oye, Frank E Harrell, Russell S Phillips, et al. 1995. A Controlled Trial to Improve Care for Seriously Ill Hospitalized Patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). *Journal of the American Medical Association* 274, 20 (1995), 1591–1598.
- [6] David R Cox. 1972. Regression Models and Life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 2 (1972), 187–202.
- [7] Chang Cui, Yongqiang Tang, and Wensheng Zhang. 2024. Deep Survival Analysis With Latent Clustering and Contrastive Learning. *IEEE Journal of Biomedical and Health Informatics* 28, 5 (May 2024), 3090–3101.
- [8] Tanujit Dey, Stuart R Lipsitz, Zara Cooper, Quoc-Dien Trinh, Martin Krzywinski, and Naomi Altman. 2022. Survival Analysis—Time-to-event Data and Censoring. *Nature Methods* 19, 8 (2022), 903.
- [9] Angela Dispenzieri, Jerry A Katzmman, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton, et al. 2012. Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population. In *Mayo Clinic Proceedings*, Vol. 87. Elsevier BV, 517–523.
- [10] Benjamin Elder, Matthew Arnold, Anupama Murthi, and Jiří Navrátil. 2021. Learning Prediction Intervals for Model Performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7305–7313.
- [11] Fahimeh Fakour, Ali Mosleh, and Ramin Ramezani. 2024. A Structured Review of Literature on Uncertainty in Machine Learning & Deep Learning. *arXiv preprint arXiv:2406.00332* (2024).
- [12] David Faraggi and Richard Simon. 1995. A Neural Network Model for Survival Data. *Statistics in Medicine* 14, 1 (1995), 73–82.
- [13] Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. 2016. Gaussian Processes for Survival Analysis. In *Advances in Neural Information Processing Systems*. 5021–5029.
- [14] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*. PMLR, 1050–1059.
- [15] Jakob Gawlikowski, Cedric Rovile Njéutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A Survey of Uncertainty in Deep Neural Networks. *Artificial Intelligence Review* 56 (2023), 1513–1589.
- [16] Xiao Gu, Yu Liu, Zaineb Mohsin, Jonathan Bedford, Anshul Thakur, Peter Watkinson, Lei Clifton, Tingting Zhu, and David Clifton. 2025. Are Time Series Foundation Models Ready for Vital Sign Forecasting in Healthcare?. In *Machine Learning for Health (ML4H)*. PMLR, 401–419.
- [17] FE Harrell. 1982. Evaluating the Yield of Medical Tests. *Journal of the American Medical Association* 247, 18 (1982), 2543–2546.

- [18] Wenchong He and Zhe Jiang. 2023. A Survey on Uncertainty Quantification Methods for Deep Neural Networks: An Uncertainty Source Perspective. *arXiv preprint arXiv:2302.13425* (2023).
- [19] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirmer, and Frank Hutter. 2025. Accurate Predictions on Small Data with A Tabular Foundation Model. *Nature* 637, 8045 (2025), 319–326.
- [20] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning* 110, 3 (2021), 457–506.
- [21] National Cancer Institute. 2024. SEER Incidence Data, 1975–2021. <https://seer.cancer.gov/data/>.
- [22] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. Random Survival Forests. *The Annals of Applied Statistics* 2, 3 (Sept. 2008). <https://doi.org/10.1214/08-aos169>
- [23] Vincent Jeanselme, Nikita Agarwal, and Chen Wang. 2024. Review of Language Models for Survival Analysis. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- [24] Edward L Kaplan and Paul Meier. 1958. Nonparametric Estimation From Incomplete Observations. *J. Amer. Statist. Assoc.* 53, 282 (1958), 457–481.
- [25] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology* 18, 1 (Feb. 2018).
- [26] Håvard Kvamme and Ørnulf Borgan. 2021. Continuous and Discrete-time Survival Prediction with Neural Networks. *Lifetime Data Analysis* 27, 4 (2021), 710–736.
- [27] Håvard Kvamme and Ørnulf Borgan. 2023. The Brier Score Under Administrative Censoring: Problems and A Solution. *Journal of Machine Learning Research* 24, 2 (2023), 1–26.
- [28] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* 20, 129 (2019), 1–30.
- [29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*.
- [30] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2314–2321.
- [31] Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen. 2023. Uncertainty Estimation in Deep Bayesian Survival Models. In *IEEE EMBS International Conference on Biomedical and Health Informatics*. IEEE, 1–4.
- [32] Christian Marius Lillelund, Martin Magris, and Christian Fischer Pedersen. 2024. Efficient Training of Probabilistic Neural Networks for Survival Analysis. *IEEE Journal of Biomedical and Health Informatics* (2024).
- [33] Tie-Yan Liu et al. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
- [34] Christos Louizos and Max Welling. 2017. Multiplicative Normalizing Flows for Variational Bayesian Neural Networks. In *International Conference on Machine Learning*. PMLR, 2218–2227.
- [35] Hrushikesh Loya, Pranav Poduval, Deepak Anand, Neeraj Kumar, and Amit Sethi. 2020. Uncertainty Estimation in Cancer Survival Prediction. In *ICLR Workshop on AI for Affordable Healthcare*.
- [36] Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. 2024. Uncertainty Quantification and Interpretability for Clinical Trial Approval Prediction. *Health Data Science* 4 (Jan. 2024).
- [37] Andrey Malinin and Mark Gales. 2018. Predictive Uncertainty Estimation via Prior Networks. *Advances in Neural Information Processing Systems* 31 (2018).
- [38] Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. 2018. Deep Survival Analysis: Nonparametrics and Missingness. In *Machine Learning for Healthcare Conference*. PMLR, 244–256.
- [39] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. 2021. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics* 25, 8 (Aug. 2021), 3163–3175.
- [40] Konstantin Posch, Jan Steinbrener, and Jürgen Pilz. 2019. Variational Inference to Measure Model Uncertainty in Deep Neural Networks. *arXiv preprint arXiv:1902.10189* (2019).
- [41] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. 2016. Deep Survival Analysis. In *Machine Learning for Healthcare Conference*. PMLR.
- [42] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. 2019. Deep Recurrent Survival Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4798–4805.
- [43] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems*. 3183–3193.
- [44] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. 2023. Post-hoc Uncertainty Learning Using a Dirichlet Meta-Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9772–9781.
- [45] Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. 2024. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. In *International Conference on Learning Representations*.
- [46] Zhaohong Sun, Wei Dong, Jinlong Shi, Kunlun He, and Zhengxing Huang. 2021. Attention-Based Deep Recurrent Model for Survival Prediction. *ACM Transactions on Computing for Healthcare* 2, 4 (Sept. 2021), 1–18.
- [47] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2023. Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation. *Transactions on Machine Learning Research* (2023).
- [48] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty Estimation Using A Single Deep Deterministic Neural Network. In *International Conference on Machine Learning*. PMLR, 9690–9700.
- [49] Ping Wang, Yan Li, and Chandan K Reddy. 2019. Machine Learning for Survival Analysis: A Survey. *Comput. Surveys* 51, 6 (2019), 1–36.
- [50] Zifeng Wang and Jimeng Sun. 2022. SurvTRACE: Transformers for Survival Analysis with Competing Events. In *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 1–9.
- [51] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. 2020. Hyperparameter Ensembles for Robustness and Uncertainty Quantification. *Advances in Neural Information Processing Systems* 33 (2020), 6514–6527.
- [52] Simon Wiegbe, Philipp Koppe, Raphael Sonabend, Bernd Bischl, and Andreas Bender. 2024. Deep Learning for Survival Analysis: A Review. *Artificial Intelligence Review* 57, 3 (2024), 65.
- [53] Tong Xia, Ting Dang, Jing Han, Lorena Qendro, and Cecilia Mascolo. 2024. Uncertainty-aware Health Diagnostics via Class-balanced Evidential Deep Learning. *IEEE Journal of Biomedical and Health Informatics* 28, 11 (2024), 6417–6428.
- [54] Xuanlong Yu, Gianni Franchi, Jindong Gu, and Emanuel Aldea. 2024. Discretization-Induced Dirichlet Posterior for Robust Uncertainty Quantification on Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6835–6843.
- [55] Yuwei Zhang, Tong Xia, Abhirup Ghosh, and Cecilia Mascolo. 2023. Uncertainty Quantification in Federated Learning for Heterogeneous Health Data. In *KDD Workshop on Federated Learning for Distributed Data Mining*.
- [56] Qixian Zhong, Jonas Mueller, and Jane-Ling Wang. 2021. Deep Extended Hazard Models for Survival Analysis. In *Advances in Neural Information Processing Systems*. 15111–15124.

A Experiment Setup

A.1 Survival Model Details

To quantify the uncertainty for survival models, we first pretrain several survival models to be quantified. Specifically, the implementations of DeepSurv [25] and DeepHit [30] are sourced from pycox package. The implementation of RSF is from scikit-survival package⁴, and the implementation of DSM is from auton_survival package⁵. The implementation of BNNSurv is from bnnsurv package⁶. We summarize the hyperparameter settings of such models in Table 5-8. Detailed implementation codes are provided in the anonymous link. Since our primary focus is on uncertainty quantification rather than survival prediction, we tune the hyperparameters of the survival models to match the performance levels reported in existing works.

Table 5: Hyperparameters for DeepSurv model.

Dataset	learning rate	dropout	hidden layers
FLCHAIN	0.01	0.1	[32]
SUPPORT	0.1	0.1	[32]
SEER-BC	0.01	0.1	[32]
SAC3	0.01	0.1	[32]

⁴<https://scikit-survival.readthedocs.io/en/latest/index.html>

⁵<https://autonlab.org/auton-survival/>

⁶<https://github.com/thecml/UE-BNNSurv>

Table 6: Hyperparameters for DeepHit model. d_{in} is determined by the dataset covariate dimensions.

Dataset	learning rate	dropout	hidden layers
FLCHAIN	0.001	0.1	$[3d_{in}, 5d_{in}, 3d_{in}]$
SUPPORT	0.005	0.1	$[3d_{in}, 5d_{in}, 3d_{in}]$
SEER-BC	0.001	0.1	$[3d_{in}, 5d_{in}, 3d_{in}]$
SAC3	0.001	0.1	$[3d_{in}, 5d_{in}, 3d_{in}]$

Table 7: Hyperparameters for DSM model.

Dataset	learning rate	Distribution, No.	hidden layers
FLCHAIN	0.005	Weibull, 4	$[32, 32]$
SUPPORT	0.005	Weibull, 6	$[32, 32]$
SEER-BC	0.005	Weibull, 6	$[32, 32]$
SAC3	0.005	Weibull, 10	$[32, 32, 32]$

Table 8: Hyperparameters for RSF model.

Dataset	n_estimators	min_samples_split	min_samples_leaf
FLCHAIN	100	20	5
SUPPORT	100	20	5
SEER-BC	100	20	5
SAC3	20	20	5

A.2 Baseline Details

For each testing sample, MC-Dropout quantifies uncertainty by activating dropout layers in the survival models during inference and calculating the maximum standard deviation of predicted survival probabilities across time steps over 100 forward passes. For the Ensemble method, we train the survival model with 10 different random initializations and use the maximum standard deviation of predicted survival probabilities across time steps from the 10 models as the uncertainty measure. It’s worth noting that we also explored alternative approaches, such as calculating the mean standard deviation and the mean Euclidean distance between predicted survival probability time vectors to assess the discrepancy among multiple predictions for the same sample. However, these alternatives performed relatively worse compared to using the maximum standard deviation.

B Additional Experiment Results

B.1 Survival Model Validation

In Table 9, we compare the performance of our implemented survival models with the reported performance on FLCHAIN and SUPPORT datasets from existing works. The results are sourced from papers [7, 25, 26, 28, 30, 39, 46, 56]. Due to variations in dataset splits, the reported performance in the literature is not exactly the same. To account for this, we present the available performance range for comparison.

B.2 Additional Results on Selective Prediction

Table 13 is a supplemented table for Table 2. We also report the IBS comparison in Table 14 and Figure 7. Except the results with DeepHit model, similar trends to C^{td} can be observed for IBS, which further demonstrate the superiority of SurvUnc framework.

Table 9: Performance comparison of survival models with reported results in literature (Lit.).

Model	From	FLCHAIN		SUPPORT	
		C-index	IBS	C-index	IBS
DeepHit	Lit.	0.79-0.80	0.09-0.13	0.56-0.64	0.20-0.23
	Ours	0.797	0.125	0.626	0.209
DeepSurv	Lit.	0.79-0.80	0.09-0.10	0.57-0.61	0.19-0.21
	Ours	0.798	0.101	0.606	0.195
RSF	Lit.	0.77-0.82	0.09-0.10	0.61-0.63	0.18-0.21
	Ours	0.795	0.100	0.631	0.189
DSM	Lit.	0.79-0.80	0.10-0.11	0.60-0.61	0.20-0.21
	Ours	0.799	0.102	0.610	0.197

In addition, the Brier score is known to be imperfect in survival analysis since it approximates the ground truth survival curve as a step function that begins at 1 and then immediately becomes 0 at the time to event. Thus, we compute the absolute difference between predicted median survival ($\min\{t|S(t|x) \leq 0.5\}$) and actual event times for uncensored patients in selective prediction on SEER-BC and SUPPORT, as shown in Table 10. Our proposed SurvUnc still achieves the best performance.

Table 10: Absolute difference between predicted median survival and actual event times of DeepSurv under different discarding percentages (10%, 30%, 50%), determined by the SurvUnc framework with different UQ methods across datasets.

Datasets	SUPPORT			SEER-BC		
	10%	30%	50%	10%	30%	50%
UQ Methods						
MC-Dropout	289.50	320.36	323.45	41.21	44.18	47.22
Ensemble	282.52	299.62	314.78	39.64	39.73	41.29
SurvUnc-RF	242.70	203.24	166.85	36.29	28.39	21.75
SurvUnc-MLP	243.22	203.77	169.99	36.40	27.79	21.81

B.3 Additional Results on Misprediction

We present the misprediction results of BNNSurv model in Table 11. It can be observed our proposed SurvUnc framework outperforms the inherent uncertainty estimation in Bayesian-based survival model, BNNSurv. Besides, the comparison between predicted uncertainty scores and IBSs for the DeepSurv model using MC-Dropout and Ensemble are shown in Figure 8 and Figure 9, respectively. Similar to selective prediction experiments, we also present misprediction results with the absolute difference between predicted median survival and actual event times for uncensored patients, as shown in Table 15.

B.4 Hyperparameter Study

For the impact of anchor number to the uncertainty quantification in Figure 6, we further present its impact with IBS in Figure 10(a). Moreover, the corresponding studies for SurvUnc-MLP are presented in Figure 10(b) and Figure 10(c) for C^{td} and IBS, respectively. The results demonstrate that the SurvUnc framework is robust with respect to the number of anchors, allowing for flexible selection.

Table 11: Misprediction detection results with BNNSurv, i.e., Pearson correlation coefficient between uncertainty scores and IBs of samples.

Datasets	FLCHAIN	SUPPORT	SEER-BC	SAC3
Bayesian	-0.426	0.321	-0.680	0.254
SurvUnc-RF	0.440	0.776	0.669	0.542
SurvUnc-MLP	0.410	0.749	0.629	0.619

B.5 Case Study

To validate the robustness of the proposed SurvUnc framework, we also evaluate two state-of-the-art models TabPFN [19] and TabNet [3] as meta models with selective prediction, as shown in Table 12. While TabPFN achieves marginally better results, the improvements

are minimal considering the substantial computational and implementation overhead.

Table 12: C^{td} of DeepSurv under different discarding percentages (10%, 30%, 50%), determined by the SurvUnc framework with different meta models across datasets.

Datasets	SUPPORT			SEER-BC		
Meta Models	10%	30%	50%	10%	30%	50%
RF	0.635	0.690	0.757	0.904	0.938	0.961
MLP	0.637	0.695	0.762	0.904	0.938	0.961
TabNet	0.639	0.695	0.762	0.904	0.938	0.961
TabPFN	0.638	0.701	0.768	0.903	0.934	0.959

Table 13: C^{td} of survival models under different discarding percentages (10%, 30%, 50%), determined by different uncertainty quantification methods across datasets. The best results are in bold. The last row of each group shows relative improvement compared with the best baseline by 100 runs of experiments. p -value < 0.001 is true for all results without *.

Survival Models	UQ Methods	FLCHAIN			SUPPORT			SEER-BC			SAC3		
		10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
DeepSurv	MC-Dropout	0.782	0.745	0.685	0.600	0.576	0.563	0.857	0.846	0.832	0.772	0.778	0.793
		±0.048	±0.061	±0.081	±0.038	±0.040	±0.048	±0.040	±0.051	±0.065	±0.034	±0.040	±0.042
	Ensemble	0.786	0.755	0.710	0.607	0.598	0.576	0.863	0.857	0.842	0.783	0.795	0.807
		±0.047	±0.059	±0.074	±0.038	±0.040	±0.048	±0.041	±0.050	±0.067	±0.033	±0.035	±0.040
	SurvUnc-RF	0.856	0.907	0.941	0.635	0.690	0.757	0.904	0.938	0.961	0.792	0.822	0.855
		±0.036	±0.028	±0.020	±0.036	±0.040	±0.039	±0.028	±0.020	±0.015	±0.031	±0.032	±0.035
	SurvUnc-MLP	0.839	0.894	0.935	0.637	0.695	0.762	0.904	0.938	0.961	0.797	0.830	0.862
		±0.042	±0.034	±0.026	±0.037	±0.041	±0.041	±0.029	±0.020	±0.014	±0.032	±0.033	±0.030
	Improv.	8.9%	20.1%	32.5%	4.9%	16.2%	32.3%	4.8%	9.5%	14.1%	1.8%	4.4%	6.8%
DeepHit	MC-Dropout	0.784	0.762	0.733	0.648	0.657	0.656	0.908	0.937	0.953	0.797	0.811	0.836
		±0.046	±0.057	±0.075	±0.033	±0.040	±0.046	±0.028	±0.019	±0.018	±0.035	±0.036	±0.034
	Ensemble	0.855	0.902	0.934	0.637	0.637	0.637	0.896	0.923	0.944	0.805	0.822	0.842
		±0.035	±0.027	±0.022	±0.034	±0.038	±0.047	±0.033	±0.029	±0.023	±0.033	±0.030	±0.034
	SurvUnc-RF	0.855*	0.908	0.940	0.639	0.664*	0.693	0.913	0.945	0.961	0.811	0.838	0.866
		±0.036	±0.027	±0.020	±0.032	±0.041	±0.047	±0.027	±0.018	±0.015	±0.031	±0.035	±0.036
	SurvUnc-MLP	0.855*	0.909	0.940	0.639	0.660*	0.692	0.914	0.945	0.961	0.817	0.845	0.876
		±0.035	±0.027	±0.020	±0.033	±0.040	±0.045	±0.027	±0.018	±0.014	±0.031	±0.031	±0.031
	Improv.	0.0%	0.8%	0.6%	-1.4%	1.1%	5.6%	0.7%	0.9%	0.8%	1.5%	2.8%	4.0%
DSM	MC-Dropout	0.780	0.741	0.687	0.613	0.614	0.613	0.866	0.854	0.832	0.793	0.776	0.770
		±0.049	±0.062	±0.079	±0.035	±0.041	±0.049	±0.037	±0.045	±0.057	±0.033	±0.040	±0.051
	Ensemble	0.787	0.755	0.715	0.616	0.605	0.581	0.872	0.868	0.854	0.802	0.813	0.828
		±0.047	±0.059	±0.075	±0.036	±0.044	±0.054	±0.036	±0.043	±0.058	±0.034	±0.037	±0.035
	SurvUnc-RF	0.854	0.907	0.941	0.640	0.685	0.725	0.910	0.943	0.960	0.813	0.837	0.862
		±0.038	±0.026	±0.020	±0.035	±0.041	±0.044	±0.027	±0.017	±0.013	±0.030	±0.033	±0.035
	SurvUnc-MLP	0.852	0.906	0.940	0.641	0.685	0.731	0.910	0.943	0.959	0.818	0.844	0.871
		±0.038	±0.028	±0.020	±0.034	±0.039	±0.043	±0.027	±0.018	±0.014	±0.029	±0.031	±0.031
	Improv.	8.5%	20.1%	31.6%	4.1%	11.6%	19.2%	4.4%	8.6%	12.4%	2.0%	3.8%	5.2%
RSF	MC-Dropout	-	-	-	-	-	-	-	-	-	-	-	-
		0.790	0.777	0.745	0.648	0.662	0.684	0.878	0.874	0.863	0.649	0.663	0.677
	Ensemble	±0.049	±0.062	±0.079	±0.037	±0.043	±0.043	±0.034	±0.042	±0.057	±0.043	±0.045	±0.050
		0.840	0.897	0.933	0.663	0.708	0.750	0.908	0.941	0.959	0.658	0.692	0.738
	SurvUnc-RF	±0.044	±0.029	±0.023	±0.037	±0.037	±0.041	±0.026	±0.019	±0.015	±0.044	±0.048	±0.053
		0.820	0.854	0.892	0.656	0.689	0.721	0.913	0.943	0.960	0.653	0.685	0.731
	SurvUnc-MLP	±0.044	±0.043	±0.041	±0.036	±0.038	±0.040	±0.026	±0.019	±0.014	±0.045	±0.050	±0.053
	Improv.	6.3%	15.4%	25.2%	2.3%	6.9%	9.6%	4.0%	7.9%	11.2%	1.4%	4.4%	9.0%
BNNSurv	Bayesian	0.773	0.732	0.670	0.623	0.648	0.687	0.847	0.836	0.805	0.719	0.740	0.771
		±0.049	±0.063	±0.082	±0.036	±0.039	±0.044	±0.044	±0.055	±0.076	±0.037	±0.040	±0.043
	SurvUnc-RF	0.848	0.902	0.936	0.650	0.701	0.764	0.891	0.932	0.955	0.727	0.765	0.808
		±0.039	±0.029	±0.023	±0.033	±0.037	±0.038	±0.034	±0.022	±0.016	±0.036	±0.034	±0.037
	SurvUnc-MLP	0.843	0.899	0.934	0.648	0.699	0.760	0.893	0.932	0.955	0.727	0.767	0.815
		±0.040	±0.030	±0.023	±0.034	±0.037	±0.038	±0.033	±0.023	±0.016	±0.036	±0.037	±0.038
		Improv.	9.7%	23.2%	39.7%	4.3%	8.2%	11.2%	5.4%	11.5%	18.6%	1.1%	3.6%

Table 14: IBS of survival models under different discarding percentages (10%, 30%, 50%), determined by different uncertainty quantification methods across datasets. The best results are in bold. The last row of each group shows relative improvement compared with the best baseline by 100 runs of experiments. p -value < 0.001 is true for all results without *.

Survival Models	UQ Methods	FLCHAIN			SUPPORT			SEER-BC			SAC3		
		10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
DeepSurv	MC-Dropout	0.268	0.302	0.344	0.151	0.164	0.169	0.240	0.259	0.282	0.120	0.123	0.122
		± 0.038	± 0.046	± 0.057	± 0.010	± 0.010	± 0.009	± 0.042	± 0.047	± 0.057	± 0.014	± 0.016	± 0.017
	Ensemble	0.265	0.289	0.318	0.147	0.154	0.161	0.232	0.242	0.257	0.115	0.115	0.119
		± 0.037	± 0.043	± 0.051	± 0.009	± 0.009	± 0.010	± 0.040	± 0.048	± 0.058	± 0.014	± 0.017	± 0.022
	SurvUnc-RF	0.219	0.184	0.152	0.134	0.118	0.100	0.200	0.160	0.134	0.104	0.090	0.080
	SurvUnc-MLP	± 0.032	± 0.033	± 0.029	± 0.009	± 0.010	± 0.012	± 0.038	± 0.029	± 0.027	± 0.013	± 0.012	± 0.014
		0.226	0.189	0.161	0.134	0.118	0.100	0.198	0.159	0.134	0.098	0.077	0.062
		± 0.034	± 0.033	± 0.032	± 0.009	± 0.011	± 0.012	± 0.039	± 0.030	± 0.026	± 0.012	± 0.011	± 0.011
	Improv.	17.4%	36.3%	52.2%	8.8%	23.4%	37.9%	14.7%	34.3%	47.9%	14.8%	33.0%	47.9%
DeepHit	MC-Dropout	0.300	0.295	0.305	0.158	0.153	0.150	0.279	0.262	0.240	0.123	0.125	0.125
		± 0.039	± 0.046	± 0.057	± 0.005	± 0.005	± 0.006	± 0.030	± 0.029	± 0.029	± 0.008	± 0.008	± 0.009
	Ensemble	0.294	0.298	0.300	0.158	0.151	0.146	0.280	0.262	0.242	0.119	0.120	0.120
		± 0.033	± 0.034	± 0.038	± 0.005	± 0.005	± 0.006	± 0.030	± 0.030	± 0.031	± 0.008	± 0.009	± 0.009
	SurvUnc-RF	0.296*	0.298*	0.303*	0.159	0.152*	0.143	0.278*	0.269	0.271	0.115	0.109	0.102
	SurvUnc-MLP	± 0.034	± 0.038	± 0.044	± 0.005	± 0.006	± 0.007	± 0.029	± 0.025	± 0.028	± 0.008	± 0.009	± 0.010
		0.299*	0.300*	0.303*	0.160	0.152*	0.145	0.276*	0.267	0.269	0.112	0.101	0.092
		± 0.035	± 0.039	± 0.040	± 0.005	± 0.006	± 0.007	± 0.029	± 0.025	± 0.026	± 0.008	± 0.009	± 0.010
	Improv.	-0.7%	-1.0%	-1.0%	-0.6%	-0.7%	2.1%	1.1%	-1.9%	-12.1%	5.9%	15.8%	23.3%
DSM	MC-Dropout	0.260	0.298	0.339	0.150	0.145	0.143	0.238	0.269	0.320	0.124	0.139	0.147
		± 0.038	± 0.046	± 0.054	± 0.009	± 0.010	± 0.011	± 0.046	± 0.058	± 0.070	± 0.017	± 0.018	± 0.019
	Ensemble	0.255	0.280	0.312	0.147	0.155	0.161	0.223	0.230	0.243	0.116	0.117	0.120
		± 0.035	± 0.042	± 0.056	± 0.010	± 0.012	± 0.014	± 0.045	± 0.054	± 0.068	± 0.017	± 0.020	± 0.022
	SurvUnc-RF	0.208	0.172	0.143	0.135	0.118	0.103	0.190	0.147	0.132	0.104	0.090	0.081
	SurvUnc-MLP	± 0.030	± 0.027	± 0.023	± 0.010	± 0.011	± 0.013	± 0.041	± 0.030	± 0.028	± 0.015	± 0.016	± 0.017
		0.211	0.175	0.144	0.135	0.120	0.107	0.190	0.149	0.132	0.095	0.075	0.064
		± 0.032	± 0.029	± 0.024	± 0.010	± 0.011	± 0.013	± 0.040	± 0.029	± 0.027	± 0.014	± 0.012	± 0.012
	Improv.	18.4%	38.6%	54.2%	8.2%	18.6%	28.0%	14.8%	36.1%	45.7%	18.1%	35.9%	46.7%
RSF	MC-Dropout	-	-	-	-	-	-	-	-	-	-	-	-
	Ensemble	0.254	0.261	0.281	0.145	0.136	0.130	0.217	0.225	0.243	0.144	0.142	0.138
		± 0.039	± 0.046	± 0.053	± 0.011	± 0.012	± 0.014	± 0.040	± 0.050	± 0.062	± 0.010	± 0.012	± 0.013
	SurvUnc-RF	0.220	0.178	0.145	0.136	0.121	0.111	0.189	0.152	0.127	0.139	0.128	0.118
		± 0.037	± 0.032	± 0.026	± 0.010	± 0.010	± 0.013	± 0.036	± 0.030	± 0.025	± 0.010	± 0.011	± 0.012
	SurvUnc-MLP	0.239	0.216	0.191	0.139	0.127	0.120	0.180	0.146	0.127	0.141	0.131	0.120
		± 0.038	± 0.042	± 0.044	± 0.010	± 0.012	± 0.013	± 0.037	± 0.027	± 0.025	± 0.011	± 0.011	± 0.012
	Improv.	13.4%	31.8%	48.4%	6.2%	11.0%	14.6%	17.1%	35.1%	47.7%	3.5%	9.9%	14.5%
BNNSurv	Bayesian	0.327	0.348	0.375	0.144	0.137	0.127	0.268	0.297	0.345	0.150	0.144	0.133
		± 0.044	± 0.054	± 0.065	± 0.009	± 0.010	± 0.013	± 0.047	± 0.059	± 0.080	± 0.015	± 0.018	± 0.023
	SurvUnc-RF	0.289	0.266	0.253	0.133	0.113	0.095	0.229	0.179	0.154	0.139	0.121	0.105
		± 0.038	± 0.038	± 0.036	± 0.008	± 0.009	± 0.009	± 0.043	± 0.032	± 0.029	± 0.013	± 0.012	± 0.013
	SurvUnc-MLP	0.292	0.268	0.251	0.133	0.114	0.095	0.225	0.178	0.152	0.137	0.114	0.095
		± 0.040	± 0.036	± 0.035	± 0.008	± 0.009	± 0.010	± 0.042	± 0.032	± 0.027	± 0.013	± 0.011	± 0.012
	Improv.	11.6%	23.6%	33.1%	7.6%	0.175%	25.2%	16.0%	40.1%	55.9%	8.7%	20.8%	28.6%

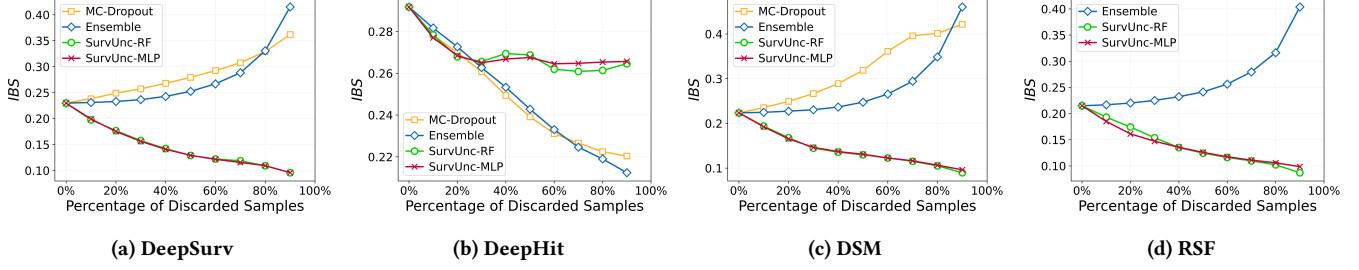


Figure 7: IBS of four survival models of (a) DeepSurv, (b) DeepHit, (c) DSM and (d) RSF on SEER-BC dataset with different percentages of samples discarded according to uncertainty scores from different uncertainty quantification methods.

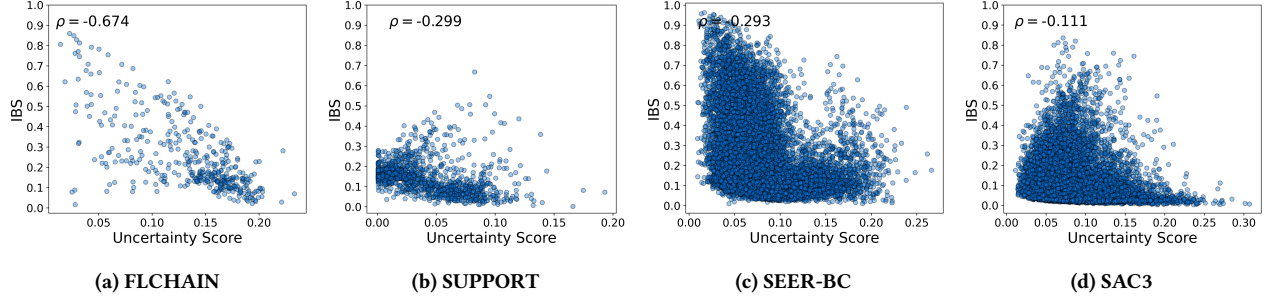


Figure 8: Predicted uncertainty scores versus IBSs from DeepSurv quantified by MC-Dropout across samples on (a) FLCHAIN, (b) SUPPORT, (c) SEER-BC and (d) SAC3 datasets.

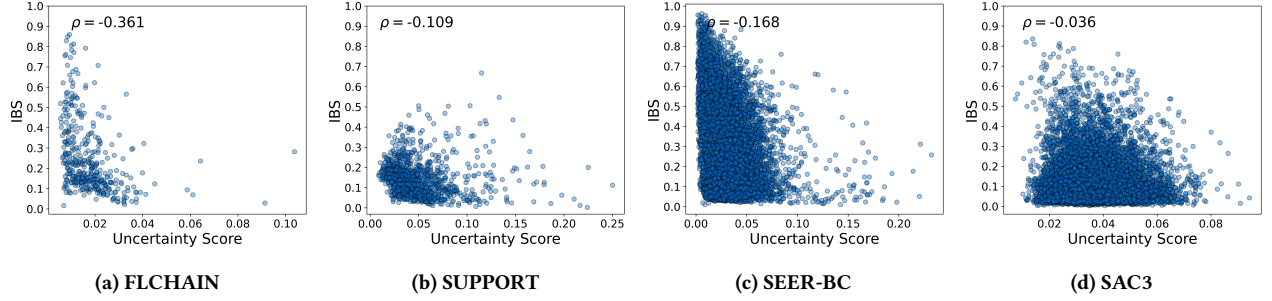


Figure 9: Predicted uncertainty scores versus IBSs from DeepSurv quantified by Ensemble across samples on (a) FLCHAIN, (b) SUPPORT, (c) SEER-BC and (d) SAC3 datasets.

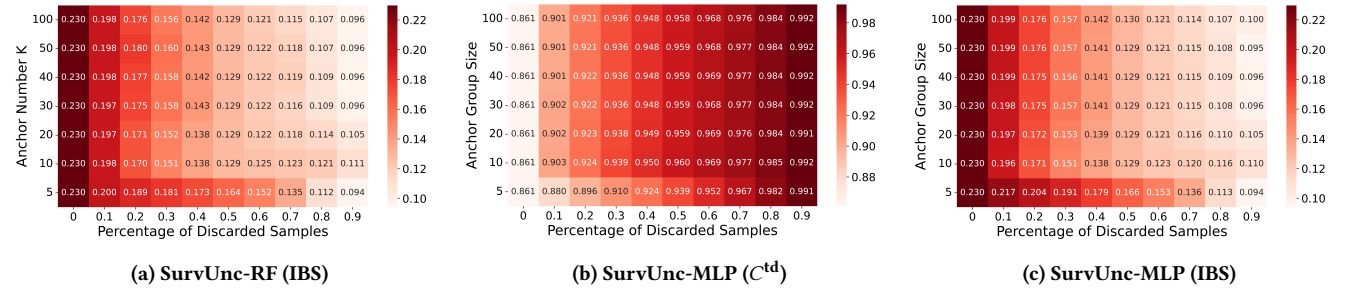


Figure 10: Varying anchor number K on selective prediction performance (a) IBS: SurvUnc-RF, (b) C^{td} : SurvUnc-MLP and (c) IBS: SurvUnc-MLP with DeepSurv on SEER-BC.

Table 15: Misprediction detection results, i.e., Pearson correlation coefficient between uncertainty scores and the absolute difference between predicted median survival and actual event times of samples.

Datasets	SUPPORT				SEER-BC			
UQ Methods	DeepSurv	DeepHit	DSM	RSF	DeepSurv	DeepHit	DSM	RSF
MC-Dropout	-0.136	-0.097	-0.008	-	-0.245	0.209	-0.677	-
Ensemble	-0.030	-0.130	-0.209	0.150	-0.076	0.267	-0.079	-0.120
SurvUnc-RF	0.479	0.028	0.415	0.503	0.643	0.027	0.626	0.536
SurvUnc-MLP	0.450	0.012	0.370	0.373	0.647	0.032	0.625	0.586