POLITECNICO DI TORINO
Repository ISTITUZIONALE

Bridging the gap: improve neural survival models with interpolation techniques

Original
Bridging the gap: improve neural survival models with interpolation techniques / Archetti, A., Stranieri, F., Matteucci, M.. - In: PROGRESS IN ARTIFICIAL INTELLIGENCE. - ISSN 2192-6352. - (2024). [10.1007/s13748-024-00343-y]

Availability:
This version is available at: 11583/2992784 since: 2024-09-27T08:55:07Z

Publisher:
Springer

Published
DOI:10.1007/s13748-024-00343-y

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s13748-024-00343-y

(Article begins on next page)

16 November 2026

Bridging the Gap: Improve Neural Survival Models with Interpolation Techniques

Alberto Archetti[1,3*], Francesco Stranieri[2,3] and Matteo Matteucci[1]

[1]Politecnico di Milano, Via Ponzio 34/5, Milan, 20133, Italy.
[2]Università degli Studi di Milano-Bicocca, Viale Sarca 336, Milan, 20126, Italy.
[3]Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy.

*Corresponding author(s). E-mail(s): alberto.archetti@polimi.it;
Contributing authors: francesco.stranieri@polito.it; matteo.matteucci@polimi.it;

Abstract

Survival analysis is an essential tool in healthcare for risk assessment, assisting clinicians in their evaluation and decision making processes. Therefore, the importance of using expressive and high-performing survival models is crucial. With the advent of neural networks and deep learning, a new generation of survival models has emerged, offering state-of-the-art capabilities to capture the non-linear and complex relationships inherent in multimodal patient data for survival prediction. However, these models often produce discrete outputs, resulting in survival functions that are coarse-grained and difficult to interpret. This study advances previous research by further exploring interpolation techniques as a post-processing strategy to improve the predictive accuracy of survival models. Our results show how the use of specific interpolation techniques significantly improves the concordance and calibration of survival estimates. This analysis encompasses a wide array of medical datasets, models, and interpolation techniques, demonstrating the effectiveness of the proposed approach and providing actionable insights for survival model design.

Keywords: survival analysis, neural networks, machine learning, interpolation techniques, healthcare

1 Introduction

Machine learning algorithms have significantly advanced the field of medical decision making and treatment planning. By harnessing the predictive power of these models, clinicians can extract meaningful insights from complex, multimodal, and nonlinear data. In particular, event time prediction plays a critical role in improving the effectiveness of clinicians' assessments and treatment strategies. In this context, survival analysis [1], a specialized branch of statistics, focuses on analyzing time-to-event data to generate event probability functions. Using patient covariate characteristics, denoted as $\mathbf{x}$, survival models compute a survival function S, represented as

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}),$$

where $S(t|\mathbf{x})$ is the probability that a subject will not experience an event up to time t. This analysis is critical for identifying specific risk factors for individual subjects, with common events in healthcare including disease onset, death, relapse, and hospital discharge. Survival analysis differs from traditional machine learning tasks such as classification and regression in that it can handle censored data points, i.e., data samples in which

the event of interest has not occurred for a given subject by the end of the observation period. Censored samples can be prevalent in several clinical datasets due to the long duration of medical trials and the privacy issues inherent in medical data collection. These issues pose significant challenges to the use of data-intensive survival models in healthcare.

Deep learning models based on neural networks are becoming increasingly popular in survival analysis applications [2]. The advantage over classical techniques lies in their ability to identify nonlinear patterns in survival data. In fact, neural network models – such as DeepSurv [3] – are more powerful than the popular linear Cox model, which is based on the assumption that the risk ratio of experiencing certain events remains constant over time across subjects [4]. While neural models rival the generalization capabilities of decision tree-based ensemble methods [5–7], they share a common limitation: the estimation of survival outcomes only at a discrete set of time points. Several methods have been developed to overcome this limitation and allow fine-grained risk assessment over time. For example, some approaches redefine survival analysis as an estimation problem over a finite set of time points [8]. While this approach naturally integrates with neural networks, it may limit the estimation accuracy of these models in real-world scenarios. Conversely, other studies have focused on time-continuous adaptation, either by refining classical survival models [3] or by modeling survival functions as piecewise constants [8, 9].

This work builds on our previous study of interpolation techniques to bridge the gap between the discrete outputs of neural network-based survival models, referred to as neural survival models, and the continuous domain of survival analysis [10]. Here, our goal is to extend the evaluation concerning the effectiveness of various interpolation methods in converting the discrete-time outputs of neural networks into continuous-time predictions. Our experimental evaluation covers a wide range of healthcare datasets with different data cardinalities and proportions of censored observations. We explore a spectrum of interpolation strategies, including linear functions, piecewise exponentials, and splines, applied without any prior knowledge of the survival data. To improve our approach and complete the analysis, we introduce a new interpolation strategy that incorporates the Kaplan-Meier estimator [11] as prior knowledge for predicting the survival function.

We evaluate the performance impact of incorporating interpolation into survival models using all the standard survival metrics: concordance index (C-Index), Integrated Brier Score (IBS), and Cumulative Area Under the Curve (AUC). Our results consistently show that interpolation techniques, regardless of their specific nature, significantly improve model performance. This improvement is particularly relevant for time-dependent metrics such as IBS and Cumulative AUC, which are more affected by fine-grained changes in the survival function. Notably, the benefit of interpolation becomes even more pronounced when the model's output neurons are significantly fewer than the number of unique events in the dataset, highlighting the critical role of interpolation in defining a more accurate survival function. These findings complete our previous analysis in [10], indicating that the application of interpolation as a post-processing technique serves as an effective regularization strategy in the design of survival models. By integrating interpolation, we can simplify the model architecture by reducing the number of output neurons, which in turn lowers the data demands for successful model training. This aspect is especially advantageous in clinical settings, where data availability is often limited – but not as limited to prevent the use of neural networks. On top of that, interpolation promotes smoother predictions, aligning with the expected real-world behavior of survival functions, thereby enhancing the model's generalization capabilities.

In conclusion, this research provides a thorough evaluation of interpolation as a lightweight post-processing technique for improving the predictive capabilities of survival models across multiple datasets, models, and evaluation metrics. By implementing this simple post-processing step, survival models can achieve improved accuracy in predicting time-to-event outcomes, thereby providing valuable support to clinicians in patient care and treatment planning.

2 Background and Related Work

This section covers the definition of survival analysis and censored data, as well as an overview of nonparametric estimation of survival functions and metrics for model evaluation. Finally, the section defines the state-of-the-art neural survival models.

2.1 What is Survival Analysis

Survival analysis is a branch of machine learning and statistics that focuses on modeling the time of occurrence of specific events within a given population. Event time estimation is particularly valuable in healthcare, where it is used to predict critical outcomes such as patient mortality, disease onset, relapse, and hospital discharge. As such, survival analysis plays a critical role in enabling medical professionals to assess the risk of patients experiencing specific events. Survival models leverage patient features, represented by a vector $\mathbf{x} \in \mathbb{R}^d$ with d dimensions, to generate a survival function defined as

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}).$$

This function describes the probability that an individual will not experience the event of interest up to time t. Each survival function starts at 1 for $t = 0$, indicating the absence of the event at the beginning of the study. The function then decreases progressively toward 0 as $t \to \infty$, ensuring event occurrence over an infinite time scale. Lastly, the survival function is non-increasing.

Certain survival models rely on estimating the instantaneous risk associated with an event over time, rather than directly calculating the probability of survival. This instantaneous risk, also known as the hazard function, is expressed as

$$h(t|\mathbf{x}) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t, \mathbf{x})}{\delta t}.$$

This function quantifies the immediate risk of the event occurring at time t, given that it has not yet occurred. Despite the focus on the hazard function in some models, it is possible to compute the survival function from the hazard itself. The relationship between the two is defined as

$$S(t|\mathbf{x}) = \exp\left(-\int_0^t h(u|\mathbf{x})\, du\right). \tag{1}$$

2.2 Censored Data

In survival experiments, it is common for some subjects not to experience an event of interest, such as failure, death, or another endpoint, within the data collection period. In such cases, a subject i is observed to survive up to a maximum observed time T_{MAX}, but the exact time of the event t_i remains undetermined. The only known information about t_i is that it exceeds T_{MAX}. Therefore, the event time is only known to be later than the subject's last follow-up. This scenario is the most common for survival applications and it is referred to as right censoring. Another common assumption on which we will rely for the rest of the work, is that censoring occurs at random. Censoring at random is a condition regarding the relationship between the censoring mechanism and the event time of interest. It is assumed that the reason an observation is censored is unrelated to the value of the observed time itself. This means the probability of an observation being censored is independent of the event time, given the observed covariates.

Censored samples populating survival datasets require careful handling when training survival models. To account for censored observations, survival datasets contain samples in the form of

$$(\mathbf{x}_i, \delta_i, t_i).$$

Here, $\mathbf{x}_i \in \mathbb{R}^d$ is the d-dimensional vector of input features for subject i, δ_i is a binary indicator reflecting whether the event was observed ($\delta_i = 1$) or censored ($\delta_i = 0$) within the data collection period, and t_i represents the observed event time or censoring time, depending the value of δ_i. This structured approach to data representation is crucial for accurately modeling survival outcomes, taking into account both uncensored and censored observations.

2.3 The Kaplan-Meier Estimator

Among the simplest yet most widely used survival models, the Kaplan-Meier (KM) estimator [11]

is known for its ability to summarize the distribution of survival outcomes within datasets. Essentially, this estimator produces an aggregate survival function that reflects the overall survival trajectory across the dataset, based solely on survival labels. As a non-parametric model, the KM estimator makes no assumptions about the probability distribution of the target data.

To compute the KM estimate for the survival function, one must identify the unique time points within the dataset, denoted as $t_1, t_2, \ldots, t_K$, for $K \leq N$ samples. At each time point t_j, the estimation requires the number of events $e_j \geq 1$ occurring at t_j and the number of subjects at risk r_j. Subjects at risk include those not yet censored or those for whom $t > t_j$. The KM estimator is then calculated as

$$S_{\mathrm{KM}}(t) = \prod_{j:t_j<t} \left(1 - \frac{e_j}{r_j}\right).$$

This formula computes the empirical survival rate at each unique time point, cumulatively multiplying these estimates to determine the overall survival probability. Notably, $S_{\mathrm{KM}}(t)$ is independent of the feature vector $\mathbf{x}$, as the KM estimator relies solely on survival labels.

2.4 Evaluation Metrics in Survival Analysis

The efficacy of survival models is evaluated using various metrics, with the C-Index, IBS, and Cumulative AUC being among the most common. The C-Index [12], in particular, serves as a measure of the predictive accuracy of survival models, quantifying the proportion of concordant pairs relative to all comparable pairs within a dataset. A pair of subjects i, j is concordant when the predicted survival order aligns with the actual event times; that is, if $t_i > t_j$, then $S(t_i|\mathbf{x}_i) \geq S(t_j|\mathbf{x}_j)$. Instead, a pair is comparable if, for $t_i < t_j$, subject i is not censored, i.e., $\delta_i = 1$. This metric estimates the probability that, for any two randomly selected subjects, the model will correctly predict the order of their survival times. The C-Index is formalized as

$$\text{C-Index} = P\left(t_i > t_j | S(t_i|\mathbf{x}_i) \geq S(t_j|\mathbf{x}_j)\right).$$

This equation evaluates the model's capacity to accurately sequence survival predictions. Notably, the C-Index does not depend on time, relying instead on pairwise comparisons across the dataset.

Due to the time-sensitive nature of survival analysis, assessing the calibration of probability estimates over time is crucial for model evaluation. The Brier score [13] is a metric specifically designed for this purpose, computing the weighted squared difference between the actual survival status of a subject i at time t and the predicted survival probability. The survival status is represented as 1 if the subject survives past time t ($t < t_i$) and 0 otherwise ($t \geq t_i$). The Brier score at time t is defined as follows:

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t)(\mathbf{1}(t < t_i) - S(t|\mathbf{x}))^2,$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, and $w_i(t)$ is a weighting factor that adjusts for the presence of censored data, thus reducing the bias from censoring in the evaluation. This adjustment is achieved through the Inverse Probability of Censoring Weighting (IPCW) method [12, 14], which assigns weights based on the inverse probability of censoring at a given time t, as

$$w_i(t) = \begin{cases} \delta_i / G(t_i) & \text{if } t_i \leq t \\ 1 / G(t_i) & \text{if } t_i > t \end{cases}. \tag{2}$$

Here, $G(t)$ represents the KM estimate of the censoring distribution, calculated over the dataset with inverted censoring indicators δ. The overall accuracy and calibration of a survival model over time can be summarized by integrating the Brier score across the entire study period, leading to the IBS:

$$\mathrm{IBS} = \frac{1}{T_{\mathrm{MAX}}} \int_0^{T_{\mathrm{MAX}}} \mathrm{BS}(t)\, dt.$$

The IBS provides a comprehensive measure of a survival model's calibration, factoring in the impact of censored data. Contrary to the C-Index, which assesses prediction ranking accuracy, the IBS evaluates the precision of survival probability estimates, where a score approaching 0 denotes optimal model calibration.

The Cumulative AUC [15] introduces a time-dependent adaptation of the traditional AUC metric, widely used in classification tasks to evaluate a classifier's ability to differentiate between classes. In the context of survival analysis, the AUC metric is tailored to assess a model's ability in distinguishing between subjects who experience the event before a specific time t and those who experience it afterward. The calculation of the AUC for survival analysis incorporates time-dependent outcomes as

$$\mathrm{AUC}(t) = \frac{\sum_i \sum_j \mathbf{1}\left(t_j > t, t_i \leq t, r_j \leq r_i\right) w_i(t)}{\left(\sum_i \mathbf{1}(t_i > t)\right)\left(\sum_i \mathbf{1}(t_i \leq t) w_i(t)\right)}.$$

In this equation, r_i and r_j denote the cumulative risk scores for two individuals i, j, and $w_i(t)$ refers to the IPCW weights of Eq. (2). By integrating the AUC over time, we obtain the Cumulative AUC, which provides a comprehensive evaluation of a model's discriminative power across the entire duration of a study. Table 1 summarizes the characteristics of the survival metrics presented.

2.5 Related Work

Neural networks have significantly improved the predictive capabilities of traditional survival analysis models. The first efforts were devoted to the extension of the Cox proportional hazard model [4] (Cox PH), a foundational framework in survival analysis, including neural networks into the computation [16]. The Cox PH model defines the hazard function as

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i^T \beta),$$

where $h_0(t)$ denotes the baseline hazard, and $\exp(\mathbf{x}_i^T \beta)$ scales this baseline hazard in accordance with individual risk factors, imposing a linear relationship between the covariates and the hazard. This formulation is based on the proportional hazard assumption, central to many classical survival analysis methods. This assumption states that the hazard ratio between any two subjects remains constant over time, enhancing model interpretability. In order to model non-linear dynamics between subject features and survival predictions, DeepSurv [3] introduced neural networks into a survival model based on the proportional hazard assumption. Like the Cox PH model, DeepSurv is trainable via the partial-log-likelihood loss function.

Nonetheless, the application of proportional hazard models to large datasets is frequently unsuccessful, as the proportional hazard assumption may not accurately represent the actual risk distribution encountered in real-life scenarios. Addressing this challenge, DeepHit [17] is as a neural survival model that utilizes sigmoid functions for estimating discrete event probabilities. Transitioning to a discretized survival analysis framework opened numerous algorithmic innovations for neural survival models. Notably, DeepHit can handle survival scenarios characterized by multiple events and competing risks through a specialized loss function that accounts for event ordering.

The Logistic Hazard model [8, 18] addresses the challenge of time-varying effects by treating survival analysis as a sequence of binary classification tasks. It predicts the risk of events at discrete time intervals using a multi-output neural network. Similarly, the Neural Multi-Task Logistic Regression (N-MTLR) model [19] extends multi-task logistic regression [20] with neural network integration, predicting event probabilities at successive time points, which are then normalized by a softmax function.

Diverging from discretization strategies, the PC-Hazard method [8, 9] assumes the hazard function to be piece-wise constant, enabling the construction of a piece-wise exponential continuous survival function. This approach employs Poisson regression for model training, allowing for the inclusion of any regression model into survival tasks. Table 2 collects the details related to the aforementioned survival models.

As a final note, this work broadens the study initiated in [10] by presenting an exhaustive analysis of interpolation techniques. This includes previously examined interpolation methods and introduces a novel interpolation approach that integrates prior knowledge through a KM estimator, offering a comprehensive exploration of methodologies.

3 Interpolation Methods

This study analyzes several interpolation methods as post-processing techniques for neural survival models. The structure of these models' outputs is

Table 1 Evaluation metrics in survival analysis and their properties.

Metrics	Optimization	Time Dependency	Target Measurement
C-Index	Max (↑)	Time independent	Risk evaluation concordance between subject pairs.
IBS	Min (↓)	Time dependent	Survival probability calibration.
Cumulative AUC	Max (↑)	Time dependent	Identify whether risk is higher or lower than a time threshold.

Table 2 Properties of survival models included in the experiments.

Model	Survival Output	Hazard Ratio	Number of Anchors	Feature Processing
Cox PH [4]	Continuous	Proportional	Fixed	Linear
DeepSurv [3]	Continuous	Proportional	Fixed	Non-Linear
DeepHit [17]	Discrete	Non-Proportional	Tunable	Non-Linear
Logistic Hazard [8, 18]	Discrete	Non-Proportional	Tunable	Non-Linear
N-MTLR [19]	Discrete	Non-Proportional	Tunable	Non-Linear
PC-Hazard [8, 9]	Continuous	Non-Proportional	Tunable	Non-Linear

defined by a series of B time instants, denoted as $\tau_1, \tau_2, \ldots, \tau_B$, where $0 < \tau_1 < \tau_2 < \cdots < \tau_B$. For a specific individual with features $\mathbf{x}$, the model generates a sequence of survival estimates $s_1, s_2, \ldots, s_B$, such that $1 \geq s_1 \geq s_2 \geq \cdots \geq s_B \geq 0$. These points, known as anchor points and represented by the pair (τ_i, s_i), form the discrete framework defining the model's survival function estimation, where $S(\tau_i|\mathbf{x}) = s_i$ for each $i = 1, \ldots, B$. To preserve the properties of the survival function, it is assumed that the points $(0, 1)$ and $(\tau_\infty, 0)$ are included as anchor points, ensuring the survival function starts at 1 and asymptotically approaches 0 over time.

Neural survival models can be categorized into proportional hazard models and non-proportional models. Within the proportional hazard category, we find DeepSurv, characterized by its neural network architecture with a single-output neuron. This neuron serves as a coefficient for modulating the baseline hazard, as denoted by the proportional hazard framework. Concerning interpolation, the most natural approach involves interpolating the baseline hazard, typically computed via the Nelson-Aalen estimator [21]. Consequently, the selection of anchor points B for interpolation is determined by the estimator, corresponding to the count of unique event times within the training dataset.

The second category covers neural survival models not relying on proportional hazard, specifically DeepHit, Logistic Hazard, and N-MTLR. In these models, the quantity of output neurons corresponds to the number of discretization bins B, set during the initial data discretization phase. This parameter B can be adjusted at the start of the experiment. As B increases, the granularity of the survival function estimation improves, enabling a more detailed analysis. However, at the same time, the parameter count of the neural network and, consequently, the model complexity increase as well. The PC-Hazard model is excluded from this discussion because, within our framework, it is equivalent to the Logistic Hazard model with the PWE interpolation technique from Section 3.3.

Next, we discuss the interpolation techniques within this framework, used to bridge the gap between anchor points. Figure 1 illustrates an example of each interpolation technique applied across all the models featured in our experiments.

3.1 Stepwise Interpolation

Stepwise interpolation is a commonly employed technique in existing literature for defining the survival function beyond the discretization points. In this approach, for any given time t within the interval $[\tau_i, \tau_{i+1})$, the survival function at t is set

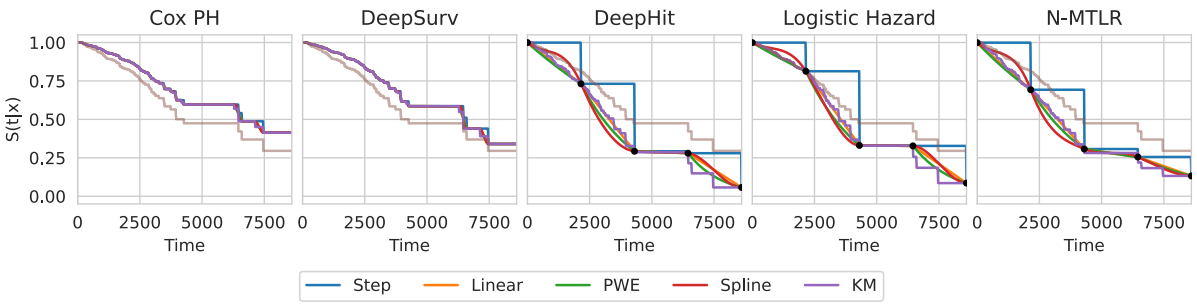

Fig. 1 Example of survival functions after interpolation for each model (columns) and interpolation method (hue) on the TCGA-BRCA [22] dataset. Each plot represents the survival function of the same subject. For non-proportional hazard models, anchor points are depicted in black. The KM estimator of each dataset is plotted in gray for reference.

to the value of the nearest preceding anchor point, τ_i, as

$$S(t|\mathbf{x}) = s_i.$$

This method results in a survival function characterized by a series of descending steps, which, while straightforward to implement – requiring only the identification of the nearest lower τ_i to time t – yields a non-continuous function. The inherent discontinuities of this function may not accurately mirror the true progression of survival functions observed in real-world scenarios, which do not typically exhibit abrupt transitions. Since this method does not effectively connect subsequent anchor points, references to stepwise interpolation and absence of interpolation will be used interchangeably throughout the rest of the work.

3.2 Linear Interpolation

A natural extension from the stepwise method is to employ linear connections between anchor points, thereby enabling the survival function to exhibit a linear transition between these points. This approach ensures function continuity and differentiability almost everywhere, except at the anchor points. The advantage of this method is mimic model smoother, gradual changes, thereby acting as a form of regularization for survival models. For a given time t within the interval $[\tau_i, \tau_{i+1})$, the survival function employing linear interpolation is calculated as

$$S(t|\mathbf{x}) = s_i + \frac{t - \tau_i}{\tau_{i+1} - \tau_i} (s_{i+1} - s_i).$$

This formula ensures a smooth transition between estimated survival probabilities, better reflecting the continuous nature of real-world survival data.

3.3 Piecewise-Exponential Interpolation

The interpolation of survival functions using piecewise exponentials comes from the PC-Hazard model [8, 9]. This model employs a neural network to estimate a series of B anchor points for the hazard function, under the assumption that the hazard function remains constant within each interval. Given the relationship between the hazard and survival functions (Eq. 1), the resulting survival function is constituted as a sequence of piecewise exponentials. The survival function at any given time t is thus calculated as

$$S(t|\mathbf{x}) = s_i \cdot \exp\left(\lambda_i \cdot \frac{t - \tau_i}{\tau_{i+1} - \tau_i}\right),$$

where the rate of change λ_i within each segment is

$$\lambda_i = \ln\left(\frac{s_{i+1}}{s_i}\right).$$

This methodology enables a continuous transition of the survival function across time, modeling an exponential pattern in the temporal dynamics of survival probability.

3.4 Monotonic Cubic Spline Interpolation

The transition to spline functions, particularly Hermit splines with a monotonicity constraint [23], represents an approach to overcoming the differentiability issue observed at the anchor points in linear and piecewise exponential interpolation methods. These methods, in fact, exhibit abrupt changes in the derivative at the anchor points. Spline functions, by contrast, offer a method for creating smooth curves through a sequence of carefully crafted polynomial functions, ensuring a continuous and differentiable non increasing survival function.

The implementation of Hermit splines for interpolating survival functions relies on the Fritsch-Carlson method, which guarantees monotonicity between a series of anchor points. The first step of this method is to compute the secant lines between consecutive anchor points as

$$\delta_i = \frac{s_{i+1} - s_i}{t_{i+1} - t_i}.$$

Then, the secant points are averaged as

$$m_i = \frac{1}{2}(\delta_{i-1} + \delta_i).$$

For the initial and final secant points, the method assumes $m_1 = \delta_1$ and $m_B = \delta_B$. Considering all other consecutive points, there are two cases. For consecutive anchors with the same value $s_i = s_{i+1}$, m_i is set to 0, as the survival function must be constant. Instead, for all the other anchor pairs, the α_i and β_i coefficients are computed as

$$\alpha_i = \frac{m_i}{\delta_i} \text{ and } \beta_i = \frac{m_{i+1}}{\delta_i}.$$

Finally, to ensure the monotonicity property, if $\alpha_i > 3$ or $\beta_i > 3$, the secant coefficient m_i are set to

$$m_i = 3\delta_i.$$

Finally, the survival function interpolated with a monotonic Hermit spline is computed as

$$\begin{aligned} S(\mu|\mathbf{x}) = \; & (2\mu^3 - 3\mu^2 + 1)s_i + \\ & (\mu^3 - 2\mu^2 + \mu)m_i + \\ & (-2\mu^3 + 3\mu^2)s_{i+1} + \\ & (\mu^3 - \mu^2)m_{i+1} \end{aligned}$$

where

$$\mu = \frac{t - t_i}{t_{i+1} - t_i}.$$

This approach accurately models continuous smooth dynamics, which are likely to represent real survival probabilities.

3.5 Kaplan-Meier Interpolation

The interpolation methods discussed thus far utilize the information of the anchor points exclusively. However, in scenarios where data are limited, adopting a smooth and symmetric interpolation approach between anchor points may not necessarily yield optimal results. To address this challenge, integrating prior knowledge about the survival dataset statistics has the potential of enhancing interpolation quality. Our novel method incorporates prior knowledge by scaling the KM estimation of survival data, $S_{\text{KM}}(t)$, between subsequent anchor points. Given two consecutive anchor points, (τ_i, s_i) and (τ_{i+1}, s_{i+1}), the idea is to squish the KM estimation in the interval $[\tau_i, \tau_{i+1})$ between the anchor values s_i and s_{i+1}. To this end, if $S_{\text{KM}}(\tau_{i+1}) \neq S_{\text{KM}}(\tau_i)$, then the survival function is computed as

$$S(t|\mathbf{x}) = s_i - \Delta_i \cdot (S_{\text{KM}}(t) - S_{\text{KM}}(\tau_{i+1}))$$

where the scaling factor Δ_i is

$$\Delta_i = \frac{s_{i+1} - s_i}{S_{\text{KM}}(\tau_{i+1}) - S_{\text{KM}}(\tau_i)}.$$

The role of Δ_i is to normalize the KM estimation between the given anchor points. Conversely, in the case where KM estimates at consecutive anchor points are equal ($S_{\text{KM}}(\tau_{i+1}) = S_{\text{KM}}(\tau_i)$), the survival function simplifies to

$$S(t|\mathbf{x}) = \frac{1}{2}(s_i + s_{i+1}),$$

which is simply the average between the given anchor points. By incorporating the KM estimation, this approach introduces a data-driven element to the interpolation, allowing for adjustments based on observed survival trends within the dataset. This method leverages the censored nature of survival data, potentially enhancing interpolation quality compared to methods that do not account for such nuances.

4 Experiments

This section collects the experimental procedures involved in the validation of interpolation techniques for neural survival models. It includes a thorough description of the datasets used in the experiments alongside a report of the architectures and parameters defined to obtain the final results. To ensure reproducibility, the source code of the experiments is publicly available [24].

4.1 Datasets

To ensure a fair evaluation and maintain consistency with similar studies, we selected our datasets from a well-known collection of benchmarks in the field of survival analysis. These datasets cover a diverse range of conditions such as breast cancer, AIDS, and cardiovascular diseases, all of which are commonly featured in survival analysis literature and are easily available through coding libraries, underscoring their broad adoption [15, 25, 26]. The selection of these datasets was intended to demonstrate the robustness and generalizability of our model across different survival distributions and contexts as in most works concerning survival algorithms rather than specific conditions [3, 5, 17].

In this study, we focused on the most common survival application: the analysis of right-censored single events. This scenario involves groups of subjects each experiencing a single event of interest, such as a disease relapse, with censoring only occurring at the end of the study. These datasets do not include the case of rare-event survival analysis, which goes beyond the scope of the work. In fact, in these settings, neural networks are generally not recommended due to their need for substantial data volumes. Thus, when the proportion of uncensored data is below 10%, more statistically solid yet less flexible methods are preferable.

The following describes the datasets utilized in the experimental analysis, detailing their content, and includes Figure 2, which illustrates the KM estimators for each dataset, providing a visual representation of survival probabilities over time. Additionally, Table 3 aggregates the summary statistics of these datasets, including censoring percentage, follow-up, and feature type distribution.

4.1.1 WHAS500

The Worcester Heart Attack Study (WHAS500) [27] is a comprehensive dataset centered around cardiovascular health, with a particular focus on individuals who have suffered from myocardial infarction, commonly known as a heart attack, between 1997 and 2001. WHAS500 tracks health outcomes of 461 patients post-myocardial infarction emplying a mix of administrative and clinical follow-up mechanisms. Features include biometric parameters, such as body-mass index, heart rate, and artial fibrillation alongside temporal features such as cohort year, lenght of hospital stay, and date of last follow-up.

4.1.2 GBSG2

The German Breast Cancer Study Group (GBSG2) [28] targets the recurrence of breast cancer post-treatment. With cancer recurrence posing a significant threat to patient recovery and survival, this dataset offers valuable insights into the effectiveness of hormone treatments and the impact of various covariates on cancer recurrence, including age, menopausal status, tumor size, and node status. The dataset originated from a randomized study conducted in Germany, which collected data on women diagnosed with breast cancer. The primary objective was to evaluate the impact of hormone therapy on cancer recurrence. The follow-up procedure was predominantly clinical, with periodic assessments to monitor patient health and detect any signs of cancer recurrence.

4.1.3 TCGA-BRCA

The Cancer Genome Atlas (TCGA) is an international genomics program aimed at characterizing and classifying genetic mutations that can lead to cancer. It includes multimodal data – tabular, image, and 3D – seeking to map the genomic changes across a wide range of cancers. Among the initiatives of TCGA, the dataset we considered focuses on the BReast CAncer study (BRCA), aiming at identifying the genetic factors that influence survival outcomes in breast-invasive carcinoma, taking into account the variability introduced by geographic regions. The study's follow-up procedures include both administrative and clinical methods. The dataset version included

Table 3 Summary statistics of the datasets used in the experiments. For each dataset, the table collects its cardinality, the percentage of censored samples, the number of numerical and categorical features, and the type of follow-up – administrative (A) or clinical (C).

Dataset	Samples	Censored	Numerical Features	Categorical Features	Follow-up
WHAS500 [27]	461	62%	7	9	A, C
GBSG2 [28]	686	56%	5	3	C
TCGA-BRCA [22]	1048	86%	1	38	A, C
METABRIC [3]	1904	42%	5	4	A, C
AIDS [29]	2839	38%	1	3	–
SUPPORT [30]	9105	32%	24	11	A, C

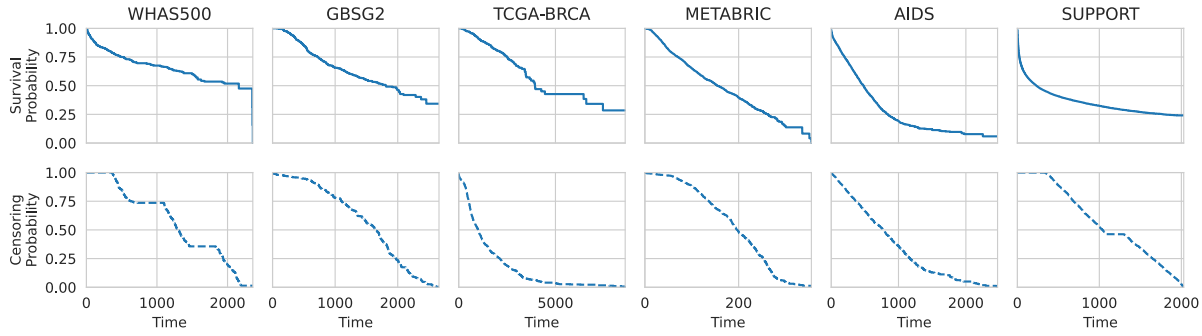

Fig. 2 Survival probability (first row) and censoring probability (second row) of the datasets included in the analysis. The survival probability is obtained with the KM estimator of the origianl dataset, while the censoring probability is the KM estimator of the dataset with inverted event labels.

in this work is obtained from Flamby [22], a collection of healthcare datasets for federated learning. To adapt the dataset to a centralized setting, we collected all data in a single node, while keeping the original geographic region of the patients.

4.1.4 METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [3, 31] aims at understanding breast cancer through molecular taxonomy, facilitating the development of personalized treatment strategies based on tumor genetic profiles. The original dataset encompasses a mix of clinical features and genomic data, including patient demographics, tumor characteristics, treatment details, and survival outcomes. Being an international data collection effort involving Canada and UK, this dataset guarantees a broad and diverse patient cohort. The follow-up procedures include both administrative and clinical approaches, with patients being monitored through medical records as well as direct clinical assessments. In this work, we employ the same feature subset of [3] to maximize comparability.

4.1.5 AIDS

The Australian AIDS Survival Dataset (AIDS) [29] aims to understand the survival patterns of patients diagnosed with AIDS in Australia before July 1, 1991. It includes a total of 2,839 entries, each equipped with features like age, state of origin, sex, date of diagnosis, and reported transmission category. To ensure patient confidentiality, this dataset has been released with a slight jitter preprocessing. No indication concerning the type of follow-up is provided in the dataset documentation.

4.1.6 SUPPORT

The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) [30], conducted between 1989 and 1997,

was a research initiative aimed at enhancing decision making and care for critically ill hospitalized patients, particularly those nearing the end of life. SUPPORT was structured into two phases: a two-year prospective observational study focused on characterizing care and treatment preferences, followed by a two-year controlled clinical trial. The type of follow-up procedures are both administrative and clinical, with patients being followed for six months post-study inclusion, with non-deceased participants being matched against a national death index up to 1997.

4.2 Experimental Procedure

This section outlines the experimental setup designed to assess the effectiveness of the proposed interpolation methods within neural survival models. Initially, a fixed random seed is utilized to allocate 80% of each dataset to a training set, with the remaining 20% constitutes the test set. Furthermore, the training set undergoes an additional division, using an 80-20% split to define a validation set. Stratification based on censorship labels is employed during each dataset division to maintain representative sample distributions. The results are averaged over 50 runs, each employing a distinct random seed, thereby altering dataset splits. To ensure consistent comparability conditions across different models, each model's training and evaluation are conducted on the same dataset splits within a given run.

Experiments encompass training and evaluation of several survival models, covering Cox PH [4], DeepSurv [3], DeepHit [17], Logistic Hazard [8, 18], and N-MTLR [19], each discussed in Section 2.5. Notably, Cox PH and DeepSurv rely to the proportional hazard assumption, which imposes the number of anchor points B for survival function estimation to be equivalent to the count of unique event times in the training dataset. Conversely, for the remaining models – DeepHit, Logistic Hazard, and N-MTLR – B is an adjustable hyperparameter. Our previous research [10] has demonstrated that the best survival performance across survival models is predominant at lower B values, where result changes in interpolation methods are noticeable. Consequently, this study conducts an exhaustive analysis by examining a range of B values from 2 to 10.

In the experimental setup, all models utilize neural networks for feature processing, with the exception of Cox PH, which employs a simple linear layer to translate features into subject-specific risks. The neural network architecture is the same across all models, consisting of a dense network with a single hidden layer. Specifically, the architecture features an input layer matching the number of subject features and a hidden layer comprising 32 nodes. For proportional hazard models, such as Cox PH and DeepSurv, the output layer contains a single output. In contrast, non-proportional hazard models have B outputs, corresponding to the number of discretization bins utilized. All layers except for the output one are followed by a ReLU activation function. To mitigate the risk of overfitting, particularly in smaller datasets, a dropout layer with a probability of 0.1 is implemented as a regularization strategy. The Adam optimizer, with a learning rate of 0.01, is employed for parameter optimization. Training is conducted over 300 epochs, adopting an early stopping rule with a 10-epoch patience. The batch size is fixed at 128.

Upon training completion, each model is equipped with interpolation techniques as detailed in Section 3, encompassing stepwise (Step), linear (Linear), piecewise exponential (PWE), monotonic cubic spline (Spline), and Kaplan-Meier (KM) methods. The performance of each model-interpolation method combination is then evaluated according to the C-Index, IBS, and Cumulative AUC, as defined in Section 2.4. For time-dependent metrics – IBS and Cumulative AUC – a discrete integral is calculated across 250 points, spanning the 20th to the 80th percentiles of each dataset's time distribution. This choice mitigates the influence of more volatile regions at the beginning and end of the survival timelines, where data may be less reliable due to scarcity.

4.3 Results

This section comments on the results obtained in the experiments. The analysis focuses on the variation of anchor points and censorship percentage of the datasets.

4.3.1 The Impact of Anchors

Figures 3, 4, and 5 present the performance metrics – C-Index, IBS, and Cumulative AUC – for

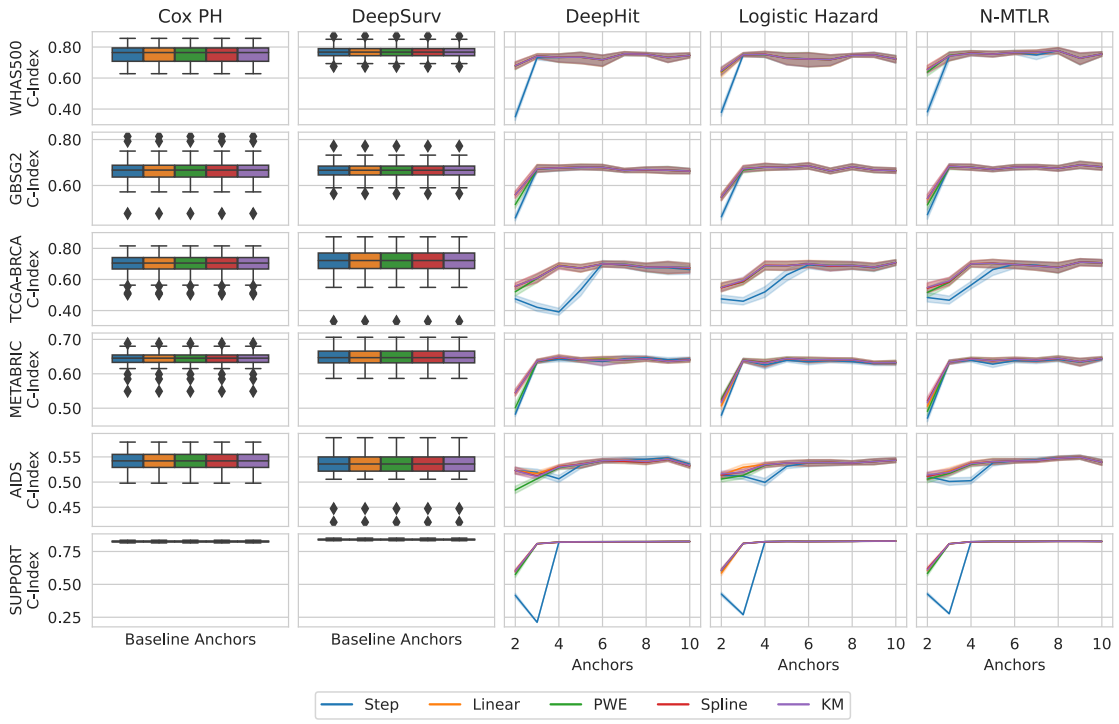

Fig. 3 C-Index (↑) for each combination of datasets (rows), models (columns), and interpolation method (hue). Anchor values are represented on the x-axis. Results are averaged over 50 runs.

the evaluated models across each datasets and the proposed interpolation methods. This analysis specifically examines the impact of varying the number of anchor points, B, within non-proportional hazard models. In fact, as already observed, for Cox PH and DeepSurv, the number of anchor points is inherently fixed, corresponding to the number of anchors of the baseline hazard function. Conversely, for the remaining models, B is a modifiable parameter, offering a tunable dimension to explore.

Analysis of Figure 3 highlights several trends across different experimental setups regarding concordance. Firstly, it is evident that the concordance index is influenced by the number of anchor points, B. Specifically, models employing Step interpolation exhibit worse concordance at lower B values ($B \leq 5$) than other interpolation methods, with performance converging towards the same value as B increases, as shown in [10]. This phenomenon aligns with expectations, as a higher B reduces the impact of interpolation by providing a more detailed survival estimation through the anchors themselves. Moreover, models based on proportional hazards demonstrate negligible variation across interpolation methods. This lack of performance difference is attributed to their fixed number of anchor points, which corresponds to the number of unique time points. Additionally, Linear, PWE, Spline, and KM interpolations show comparable concordance across the whole spectrum of anchor values tested. Notably, despite incorporating prior knowledge, KM interpolation does not consistently outperform simpler methods such as Linear in most scenarios. Remarkably, for $B \geq 4$, the concordance index across all models approximates a similar value, indicating a convergence in performance with an adequate number of anchor points.

Figure 4 showcases the IBS results, where most insights from the concordance analysis (Figure 3) remain applicable. Again, proportional hazard models exhibit minimal variation in IBS outcomes. At the same time, for non-proportional hazard models, the difference in IBS between Step interpolation and alternative methods diminishes with an increase in the number of anchor points. Furthermore, IBS results tend to converge across all

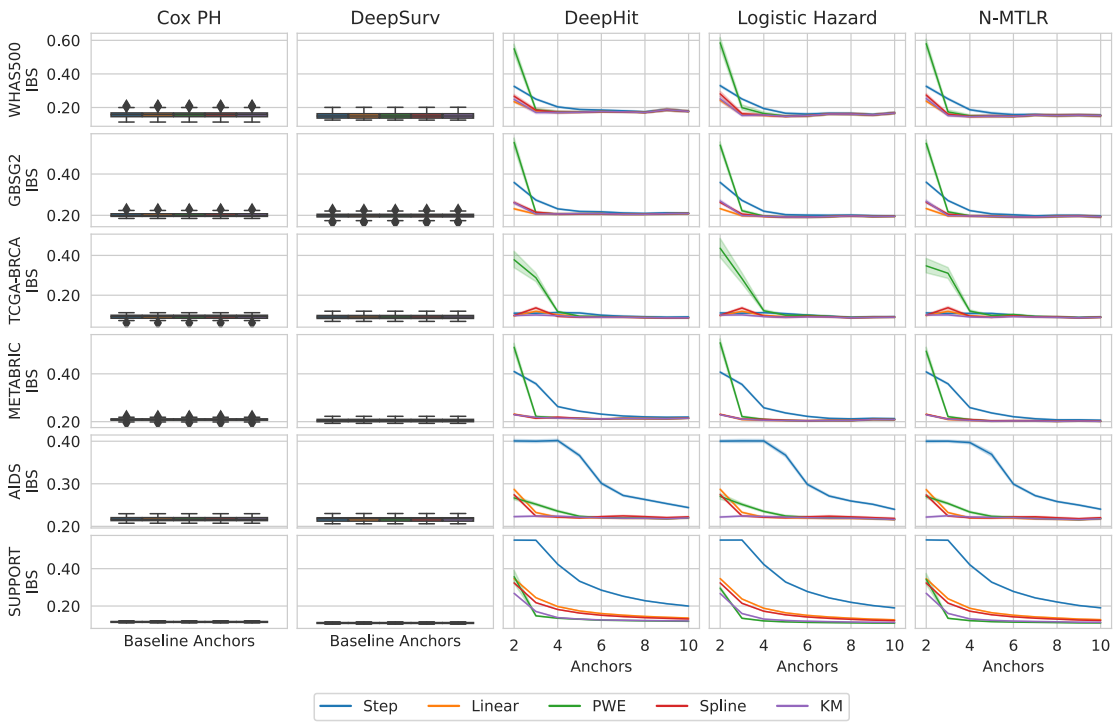

Fig. 4 IBS (↓) for each combination of datasets (rows), models (columns), and interpolation method (hue). Anchor values are represented on the x-axis. Results are averaged over 50 runs.

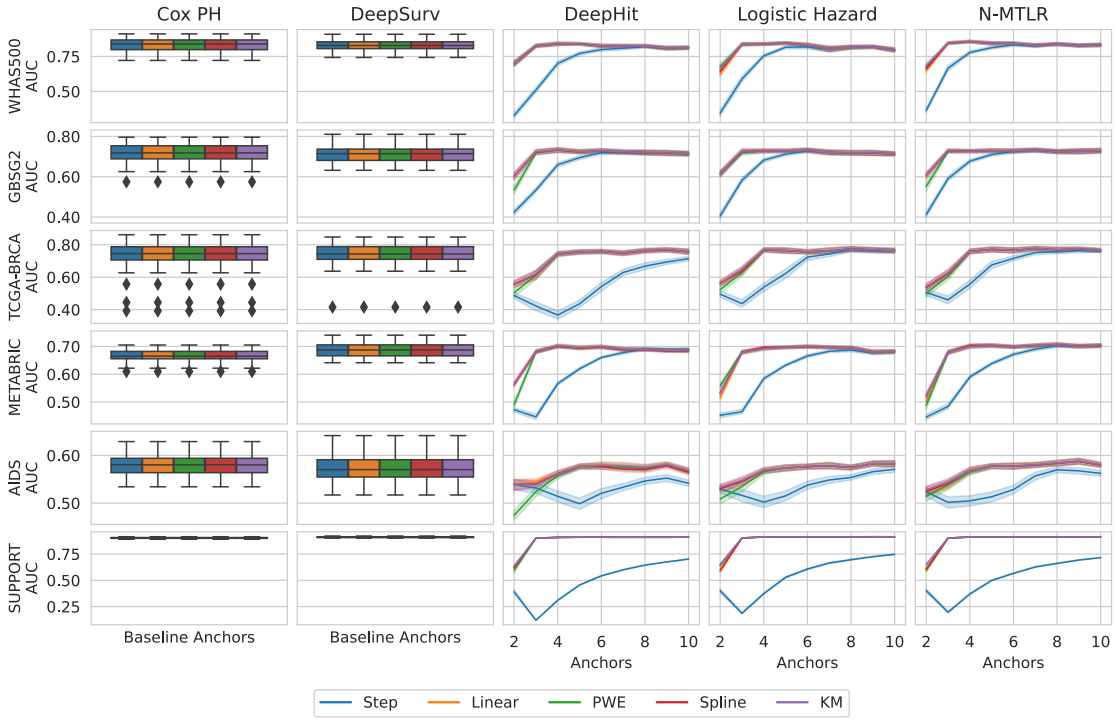

Fig. 5 Cumulative AUC (↑) for each combination of datasets (rows), models (columns), and interpolation method (hue). Anchor values are represented on the x-axis. Results are averaged over 50 runs.

models for $B \geq 4$, provided that an interpolation method other than Step is used. A distinct trend observed in the IBS analysis, diverging from prior observations, pertains to the performance of PWE interpolation. In WHAS500, GBSG2, TCGA-BRCA, and METABRIC, PWE interpolation significantly underperforms with respect to other methods, including Step, particularly at lower anchor values. On top of that, a relevant finding is the advantage of KM interpolation in terms of IBS for the largest datasets, AIDS and SUPPORT, especially in the lower anchor range ($B \leq 4$). This finding shows that the KM method, despite not improving concordance, is able to capitalize on the censoring information of the data to improve calibration. The KM advantage is noticeable in the IBS comparison as this is the metric most sensible to the censoring distribution. In particular, IPC weighting, as detailed in Section 2.4, affects the IBS importance over time, guaranteeing an advantage of KM in larger datasets, where survival dynamics between anchors are likely to be less noisy.

Finally, Figure 5 provides insights into the Cumulative AUC trends, reinforcing observations made from the analysis of Figures 3 and 4. Specifically, for non-proportional models where the number of anchor points, B, can be adjusted, a significant difference in AUC between the Step interpolation method and others is evident in lower values of B. For datasets with a smaller number of samples, WHAS500 and GBSG2, the performance difference becomes negligible after a threshold of 5-6 anchors. In contrast, for larger datasets, TCGA-BRCA, METABRIC, and AIDS, this threshold shifts up to 10. The SUPPORT dataset shows an even higher threshold, where the difference in AUC between Step and other interpolation methods remains significant beyond 10 anchors. Notably, across the analyses, KM interpolation does not exhibit advantages in terms of Cumulative AUC when compared to methods other than Step.

4.3.2 The Impact of Censoring

The analysis of interpolation results can be approached from an alternative perspective by examining the influence of the censorship percentage in survival data. Specifically, Figures 6, 7, and 8 illustrate the C-Index, IBS, and Cumulative AUC metrics for various interpolation methods across multiple dataset splits. In these experiments, a censoring threshold $c \in [0.1, 0.9]$ is set, and a number of dataset samples is randomly selected to maintain a c ratio between censored and non-censored samples. These plots display metrics for c values ranging from 10% to 90%, in increments of 2%. It is important to observe that selecting random dataset subsets based on a predefined censoring ratio may result in datasets of significantly reduced size or datasets that do not accurately represent the original population. Consequently, the subsequent analyses primarily focus on the comparative performance of interpolation methods at each specified c level, rather than the absolute performance metrics and their variations with c. Additionally, for each non-proportional model, the parameter B is fixed at 5. To accommodate potentially smaller datasets, the number of hidden neurons in each network is reduced from 32 to 8 and the maximum number of training epochs from 300 to 200.

Figure 6 illustrates the C-Index across varying censorship thresholds c. Building upon the observations from Section 4.3.1, data suggest that proportional hazard models are not affected by the choice of interpolation method, regardless of the censoring ratio. Conversely, in non-proportional models, slight differences in concordance can be observed between the Step interpolation method and others within TCGA-BRCA and AIDS. Notably, these differences become more pronounced at higher censoring rates.

Figure 7 shows the impact of c on the efficacy of various interpolation methods from the IBS perspective. While the observations regarding proportional hazard models remain consistent, indicating no significant influence of interpolation method on model performance, a different scenario is observed for non-proportional hazard models. Here, the choice of interpolation method markedly influences IBS. Specifically, this is the only instance where the Step method demonstrates superior performance compared to its counterparts. In fact, Step yields lower IBS values in datasets with smaller sizes, WHAS500, GBSG2, and TCGA-BRCA, at censoring ratios below 50%. Conversely, in all other scenarios, alternative interpolation techniques are more effective than Step. Particularly in larger datasets, AIDS and SUPPORT, KM outperforms Linear and Spline.

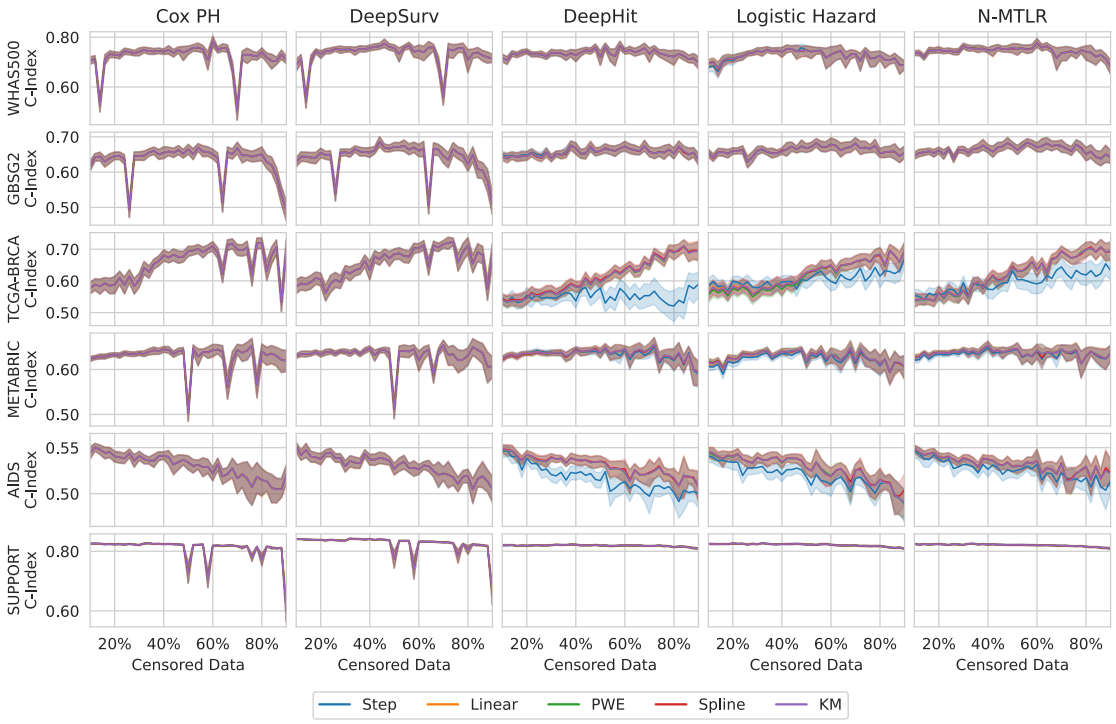

Fig. 6 C-Index (↑) for each combination of datasets (rows), models (columns), and interpolation method (hue). Censorship percentages are represented on the x-axis. Results are averaged over 50 runs.

Finally, Figure 8 collects the Cumulative AUC results across different censorship ratios, largely following the patterns observed in Figure 6. The principal distinctions in AUC are observed between Step and non-Step interpolation methods within the context of non-proportional hazard models. Notably, the discrepancy in AUC performance is more pronounced than the concordance gap observed earlier across all plots. This highlights the significant influence of interpolation method choice on model performance in terms of AUC, especially in non-proportional hazard models.

5 Discussion

Interpolation techniques serve as a postprocessing step in the implementation of neural survival models, facilitating the construction of continuous survival function estimates from a discrete set of anchor points. Our investigation into various interpolation methods, ranging from simple linear interpolation to methods integrating prior knowledge on survival data, reveals distinct trends. Notably, interpolation's impact on proportional hazard models such as Cox PH and DeepSurv is negligible. This outcome aligns with expectations, given these models' reliance on a large number of anchor points, equal to the unique time points in the training set. Therefore, the effect of interpolation is limited, due to the fine-grained survival estimation provided by the anchors themselves.

Conversely, non-proportional hazard models present a different scenario. Within the extended plethora of tested datasets, when the number of anchors B is low ($B \leq 4$), interpolation enhances survival metrics, particularly the Cumulative AUC. Even with a higher anchor count, interpolation does not degrade performance, demonstrating its potential as a beneficial or, at least, neutral addition to survival models. Interestingly, the choice of interpolation technique – be it Linear, PWE, Spline, or KM – does not significantly influence the overall outcome, provided model calibration is not the primary concern. Thus, employing the simplest Linear interpolation method can augment survival performance with minimal computational overhead.

However, in specific cases where calibration is the most critical modeling concern, the impact

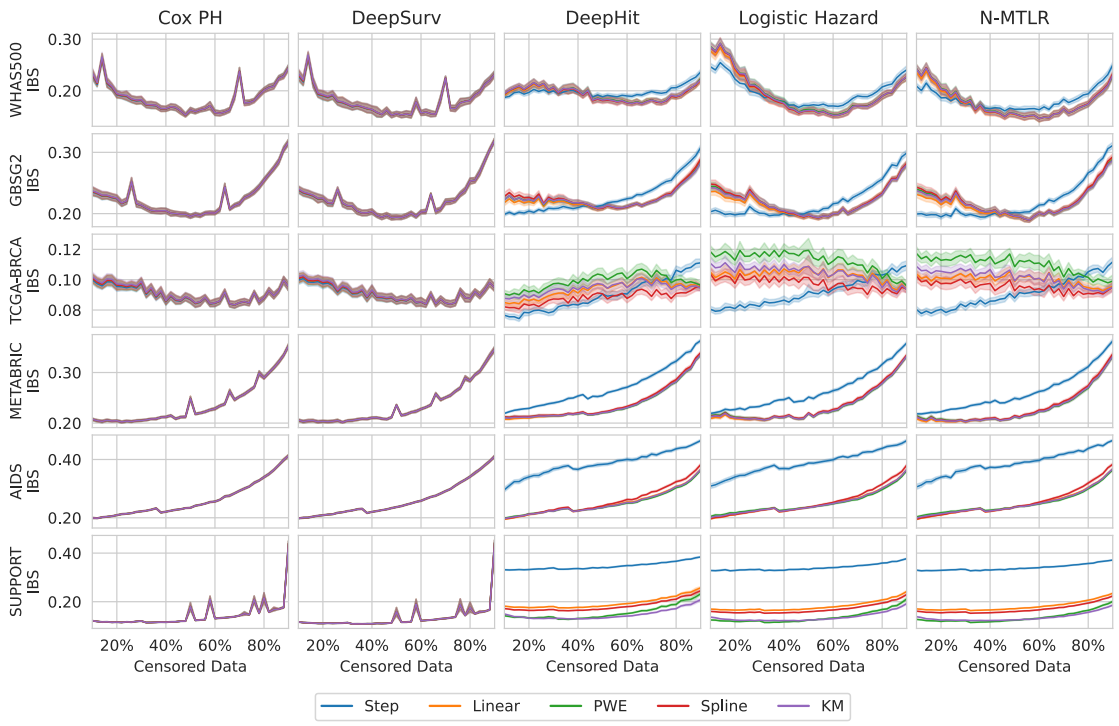

Fig. 7 IBS (↓) for each combination of datasets (rows), models (columns), and interpolation method (hue). Censorship percentages are represented on the x-axis. Results are averaged over 50 runs.

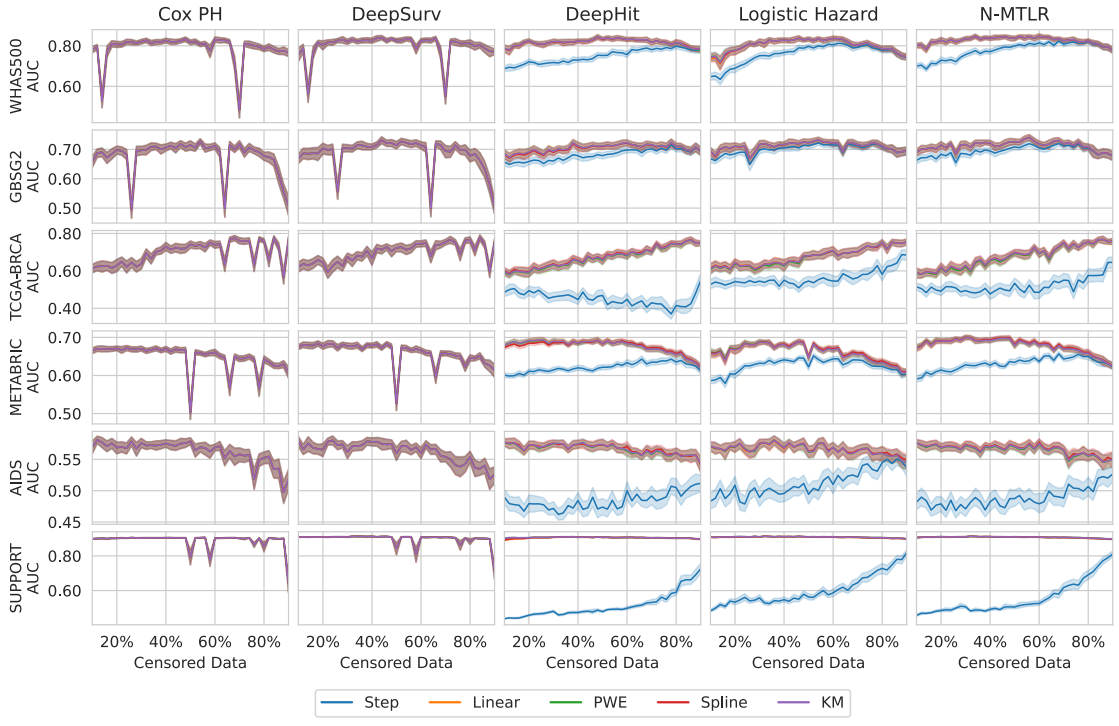

Fig. 8 Cumulative AUC (↑) for each combination of datasets (rows), models (columns), and interpolation method (hue). Censorship percentages are represented on the x-axis. Results are averaged over 50 runs.

of interpolation can differ. By inspecting the IBS metric, the naïve Step method falls short with respect to the other techniques, with few exceptions for smaller datasets and censoring ratios lower than 50%. Conversely, in most other cases, especially when censoring is higher than 50% or when datasets are very large, the KM interpolation is shown to generally outperform the others. While the difference in IBS is low, it is still statistically significant, highlighting how an interpolation technique that leverages existing censoring information in survival data can provide a slight performance boost. This difference is likely to be more significant as datasets increase, leading to the necessity of more precise interpolation methods, able to capture finer details between anchor points.

Summarizing our findings, we propose the following guidelines: for models based on proportional hazards, interpolation is unlikely to alter results significantly. For non-proportional models, instead, Linear interpolation generally enhances the overall performance. Yet, in applications demanding optimal model calibration, i.e., the IBS is the metric of primary concern, the choice of interpolation may vary, with KM interpolation being a sound choice, especially for datasets with a large cardinality or high censoring ratios.

As a final note, our research primarily aims to improve model performance in terms of survival metrics exclusively. Despite the focus on medical scenarios, it is essential to clarify that our work does not seek to draw actionable conclusions on treatment planning for specific diseases or conditions. Instead, our objective is solely to improve algorithmic modeling in survival analysis. As such, the application of these methodologies to specific medical contexts should be carefully evaluated by domain experts to ensure compatibility with the intended application scenarios.

5.1 Ethical Considerations

Our research delves into survival analysis, a field strongly related to risk assessment based on artificial intelligence within the healthcare sector. As a key statistical tool for modeling event occurrences over time, survival analysis contributes significantly to informed medical decision making and the allocation of hospital resources. However, relying solely on statistical inferences derived from deep learning algorithms may lead to overly simplistic conclusions, particularly in instances where the underlying data may be inaccurate, incomplete, or fail to reflect the complexities of real-world clinical scenarios. Consequently, we emphasize how essential is the inclusion of domain-specific expertise and experiential knowledge into the decision making process, supplementing the quantitative outputs generated by these models.

Moreover, handling patient data properly – specifically concerning transparency and consent – assumes critical importance within the healthcare domain due to the inherently sensitive nature of such information. Our investigation relies exclusively on the use of established public datasets, which are commonly utilized across the majority of studies on survival analysis research, thereby ensuring adherence to ethical standards concerning individual and data proprietor rights. Also, this approach promotes a standardized and reproducible evaluation of new algorithmic survival techniques.

In conclusion, while our study primarily concentrates on refining the algorithmic framework underlying survival models, it simultaneously acknowledges the ethical implications inherit to conducting survival analysis within a healthcare context. Our objective is to enhance the reliability of these models, advocating for a healthcare paradigm in which statistical analyses serve to support – rather than substitute – the clinical judgments of medical professionals.

6 Conclusion

In this study, we have explored the impact of various interpolation techniques on the performance of state-of-the-art neural survival models. The core of our investigation centered on the effect of interpolation in connecting the anchor points predicted by discrete neural networks, thereby allowing the evaluation of continuous survival estimates. Our empirical analysis, conducted across a comprehensive array of benchmark datasets and models, demonstrates that integrating interpolation techniques enhances both the concordance and calibration of survival models. Therefore, by bridging the gap between anchor points, interpolation not only creates smooth and realistic survival curves, but can effectively improve survival metrics with a minimal computational overhead. Data, source

code, and results of our analyses are publicly available and reproducible.

References

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

[2] Kvamme, H., Borgan, Ø., Scheel, I.: Time-to-event prediction with neural networks and cox regression. Journal of Machine Learning Research **20**, 129–112930 (2019)

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

[4] Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–220 (1972). Accessed 2023-01-11

[5] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The annals of applied statistics **2**(3), 841–860 (2008)

[6] Archetti, A., Matteucci, M.: Federated survival forests. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2023). https://doi.org/10.1109/IJCNN54540.2023.10190999

[7] Archetti, A., Ieva, F., Matteucci, M.: Scaling survival analysis in healthcare with federated survival forests: A comparative study on heart failure and breast cancer genomics. Future Generation Computer Systems **149**, 343–358 (2023) https://doi.org/10.1016/j.future.2023.07.036

[8] Kvamme, H., Borgan, Ø.: Continuous and discrete-time survival prediction with neural networks. Lifetime Data Analysis **27**(4), 710–736 (2021)

[9] Bender, A., Rügamer, D., Scheipl, F., Bischl, B.: A general machine learning framework for survival analysis. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 158–173 (2021). Springer

[10] Archetti, A., Stranieri, F., Matteucci, M.: Deep survival analysis for healthcare: An empirical study on post-processing techniques. In: Calimeri, F., Dragoni, M., Stella, F. (eds.) Proceedings of the 2nd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023). CEUR Workshop Proceedings, vol. 3578, pp. 99–121. CEUR-WS.org, Rome, Italy (2023). http://ceur-ws.org/Vol-3578/

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

[12] Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.-J.: On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine **30**(10), 1105–1117 (2011)

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

[14] Robins, J.M., Rotnitzky, A.: Recovery of information and adjustment for dependent censoring using surrogate markers. In: AIDS Epidemiology, pp. 297–331. Springer, Boston, Massachusetts, USA (1992)

[15] Pölsterl, S.: scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research **21**(212), 1–6 (2020)

[16] Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in Medicine **14**(1), 73–82 (1995) https://doi.org/10.1002/sim.4780140108 . Accessed 2022-09-01

[17] Lee, C., Zame, W., Yoon, J., Van Der Schaar,

M.: Deephit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[18] Gensheimer, M.F., Narasimhan, B.: A scalable discrete-time survival model for neural networks. PeerJ **7**, 6257 (2019)

[19] Fotso, S.: Deep neural networks for survival analysis based on a multi-task framework. arXiv preprint arXiv:1801.05512 (2018)

[20] Yu, C.-N., Greiner, R., Lin, H.-C., Baracos, V.: Learning patient-specific cancer survival distributions as a sequence of dependent regressors. Advances in neural information processing systems **24** (2011)

[21] Nelson, W.: Theory and applications of hazard plotting for censored failure data. Technometrics **14**(4), 945–966 (1972)

[22] Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., Muzellec, B., Philippenko, C., Silva, S., Teleńczuk, M., Albarqouni, S., Avestimehr, S., Bellet, A., Dieuleveut, A., Jaggi, M., Karimireddy, S.P., Lorenzi, M., Neglia, G., Tommasi, M., Andreux, M.: Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 5315–5334. Curran Associates, Inc., Long Beach, California, USA (2022)

[23] Fritsch, F.N., Carlson, R.E.: Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis **17**(2), 238–246 (1980). Accessed 2023-09-15

[24] Archetti, A.: Source Code for Interpolation in Deep Survival Analysis. GitHub. https://github.com/archettialberto/interpolation_for_deep_survival_analysis (2024)

[25] Drysdale, E.: Survset: An open-source time-to-event dataset repository. arXiv preprint arXiv:2203.03094 (2022)

[26] Archetti, A., Lomurno, E., Lattari, F., Martin, A., Matteucci, M.: Heterogeneous datasets for federated survival analysis simulation. In: Companion of the 2023 ACM/SPEC International Conference on Performance Engineering. ICPE '23 Companion, pp. 173–180. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3578245.3584935 . https://doi.org/10.1145/3578245.3584935

[27] Hosmer, D.W., Lemeshow, S., May, S.: Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (2008). https://doi.org/10.1002/9780470258019 . http://doi.wiley.com/10.1002/9780470258019 Accessed 2023-04-29

[28] Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., Rauschecker, H.: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. Journal of Clinical Oncology **12**(10), 2086–2093 (1994)

[29] Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D.: R package: MASS (2022). https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/00Index.html

[30] Knaus, W., Harrell, F., Lynn, J., Goldman, L., Phillips, R., Connors, A. Jr, Dawson, N., Fulkerson, W., Califf, R., Desbiens, N., Layde, P., Oye, R., Bellamy, P., Hakim, R., Wagner, D.: The support prognostic model. objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. Annals of internal medicine **122**, 191–203 (1995)

[31] Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A.,

Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D.W.Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J.F., Mukherjee, A., Turashvili, G., Green, A.R., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D.R., Ellis, I.O., Purushotham, A., Pinder, S.E., Børresen-Dale, A.-L., Earl, H.M., Pharoah, P.D., Ross, M.T., Aparicio, S., Caldas, C.: The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature Communications **7**(1), 11479 (2016) https://doi.org/10.1038/ncomms11479 . Accessed 2023-04-29