POLITECNICO DI TORINO
Repository ISTITUZIONALE

Deep Survival Analysis for Healthcare: An Empirical Study on Post-Processing Techniques

Original
Deep Survival Analysis for Healthcare: An Empirical Study on Post-Processing Techniques / Archetti, A., Stranieri, F., Matteucci, M.. - ELETTRONICO. - 3578:(2023), pp. 99-121. (2nd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023) Rome (Italy) November 6, 2023).

Availability:
This version is available at: 11583/2984364 since: 2023-12-06T09:32:13Z

Publisher:
CEUR Workshop Proceedings

Published
DOI:

Terms of use:

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

Publisher copyright

(Article begins on next page)

27 November 2026

Deep Survival Analysis for Healthcare: An Empirical Study on Post-Processing Techniques

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

[1]Politecnico di Milano, Via Giuseppe Ponzio, 34, 20133 Milan, Italy
[2]Università degli Studi di Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy
[3]Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10138 Turin, Italy

Abstract

Survival analysis is a crucial tool in healthcare, allowing us to understand and predict time-to-event occurrences using statistical and machine-learning techniques. As deep learning gains traction in this domain, a specific challenge emerges: neural network-based survival models often produce discrete-time outputs, with the number of discretization points being much fewer than the unique time points in the dataset, leading to potentially inaccurate survival functions. To this end, our study explores post-processing techniques for survival functions. Specifically, interpolation and smoothing can act as effective regularization, enhancing performance metrics integrated over time, such as the Integrated Brier Score and the Cumulative Area-Under-the-Curve. We employed various regularization techniques on diverse real-world healthcare datasets to validate this claim. Empirical results suggest a significant performance improvement when using these post-processing techniques, underscoring their potential as a robust enhancement for neural network-based survival models. These findings suggest that integrating the strengths of neural networks with the non-discrete nature of survival tasks can yield more accurate and reliable survival predictions in clinical scenarios.

Keywords

survival analysis, neural networks, regularization techniques, healthcare

1. Introduction

Survival analysis [1] is a field of statistics concerned with modeling time-to-event data. Its primary objective is to construct a survival function S depending on time t tailored to a particular subject, representing the probability of not experiencing a particular event of interest up to t, such as disease onset, death, or hospital discharge. Thus, a survival function is formally denoted as $S(t) = P(T > t)$. The analysis of time-to-event data is of paramount importance in healthcare, facilitating the identification of patient risk factors over time. Distinctively, survival analysis differs from conventional machine learning tasks such as classification and regression due to its ability to handle censored data points – instances where the event of interest has not yet occurred for a particular subject. This characteristic is common in clinical data, given the

HC@AIxIA 2023: Workshop on Artificial Intelligence For Healthcare, November 6, 2023, Rome, Italy
*Corresponding author.
alberto.archetti@polito.it (A. Archetti); francesco.stranieri@polito.it (F. Stranieri); matteo.matteucci@polimi.it (M. Matteucci)
0000-0003-3826-4645 (A. Archetti); 0000-0002-5366-8499 (F. Stranieri); 0000-0002-8306-6739 (M. Matteucci)
© 2023 Copyright ©2023 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3578/paper9.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

prolonged, complex, and privacy-constrained nature of data collection, which challenges the applicability of data-intensive machine learning models.

Recent advancements in survival applications exploit neural network-based deep learning techniques, emphasizing their ability to model the non-linear relationships between patient features and time-to-event records. Their utility has been demonstrated in various studies [2, 3, 4, 5, 6], emphasizing their generalization advantage over traditional statistical approaches and matching the expressive power of ensemble methods [7, 8, 9]. However, most common neural network architectures involve a set of discrete outputs, necessitating specific processing to adapt to the continuous nature of survival analysis. To this end, numerous coping strategies between discrete-output neural networks and survival analysis have been introduced. Most techniques focus on time discretization [5], enabling neural networks to encapsulate time-event associations for a limited set of time points. Instead, few methodologies directly tackle time-continuous survival functions and are based on proportional hazard [2] or piece-wise constant hazard [5, 10].

In our research, we conduct a thorough examination of multiple interpolation methods to determine if post-processing interpolation can augment the efficacy of discrete-output neural networks. Specifically, we delve into three interpolation techniques: linear, piece-wise exponential, and spline-based, applying them to the state-of-the-art neural survival models. We investigate whether performance gaps are relevant between interpolated and non-interpolated versions of the same survival model. Our investigation employs time-dependent survival metrics to gauge the efficacy of neural-based models, namely the Integrated Brier Score (IBS) and the Cumulative Area-Under-the-Curve (Cumulative AUC). Our empirical analysis, validated across several real-world healthcare datasets, indicates that interpolation supports the generalization capability of neural-based survival models. This improvement is particularly relevant when the number of discretization bins and, consequently, neural network outputs is substantially smaller than the dataset's sample count. This scenario commonly arises in practical applications where the dataset size considerably outweighs the neural network's output neurons.

In summary, our research offers a comprehensive empirical analysis of interpolation methods tailored for neural-based survival models. We explore the potential advantages of incorporating a post-processing interpolation phase based on simple operations with negligible computational overhead. These insights bear significant implications for the clinical applicability of survival models, suggesting that a simple interpolation step can markedly boost the generalizability of a neural-based survival model.

2. Background

This section provides the necessary background on survival analysis as a machine-learning problem, alongside the description of the survival metrics to assess model performance that will be investigated in subsequent experimental evaluation.

2.1. Survival Analysis

Survival analysis tackles time-to-event modeling, leveraging both statistical and machine-learning methodologies. It plays a pivotal role in interpreting clinical data, forecasting occurrences such as the onset of a disease, relapses, mortality, and hospital discharges. By harnessing patient information, the aim is to formulate a time-dependent parametric function, $S(t)$, that denotes the probability of a subject not encountering a specified event up to a given time, expressed as

$$S(t) = P(T > t). \tag{1}$$

This non-increasing function starts with a value of 1 at $t = 0$, approaching 0 as t tends to infinity.

Instead of $S(t)$, several survival methods estimate the instantaneous hazard rate for each individual, called *hazard function*:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}. \tag{2}$$

From the hazard function, the survival function can be derived as

$$S(t) = \exp(-H(t)) \tag{3}$$

where $H(t)$ represents the integral of h over the interval from 0 to t.

What sets survival analysis apart from conventional machine learning tasks, like classification or regression, is its ability to analyze censored data points. Such data represent subjects who have not encountered the specified event during the data collection period. Hence, survival datasets comprise triplets: $(\mathbf{x}_i, \delta_i, t_i)$, where (i) $\mathbf{x}_i$ indicates the feature vector for subject i; (ii) δ_i is a binary flag, which is set to 0 if the sample is censored; and (iii) t_i designates either the event's time or the censoring time, depending on the value of δ_i. This is the most common scenario in survival problems, referred to as *right censoring*. Throughout this paper, our discussions will refer to the right censoring context.

The most prevalent models used for deriving survival functions include the non-parametric Kaplan-Meier model [11] and the linear Cox model [12]. Machine learning-enhanced non-linear extensions typically employ ensemble strategies [7, 13] and neural networks, which will be analyzed in Section 3.

2.2. Metrics in Survival Analysis

The most common metrics used to evaluate the predictive power of survival models are the Concordance Index (C-Index), the IBS, and the Cumulative AUC. The C-Index [14] measures the agreement between the predicted survival outcomes from a model and the actual observed outcomes for pairs of samples. Specifically, for each time point, the predicted outcome is determined by the model's survival probability or risk score, while the true outcome reflects the event status – 1 for non-censored and 0 for censored samples. Only pairs with times $t_1 < t_2$ and events e_1, e_2 where e_1 is non-censored are considered comparable. The C-Index measures the proportion of comparable pairs that are concordant, meaning the sample with the higher predicted survival probability outlives the other. This measure can be interpreted as

the probability that, for two randomly chosen individuals, the one with the higher risk score will experience the event first. A C-Index value of 0.5 signifies random predictions, whereas 1 indicates perfect concordance. While easy to interpret, the C-Index does not provide information about model calibration.

Alongside the C-Index, another common metric for assessing survival models is the Brier Score (BS) [15], which quantifies both precision and calibration of predicted survival outcomes. The BS computes the squared difference between the actual event occurrence (1 for the event and 0 otherwise) and the predicted survival probability for a specific time instant. Ideally, a BS value should be close to 0, indicating perfect prediction. The IBS integrates the Brier scores over various times, giving an overall temporal performance evaluation of the model. The IBS summarizes the model's ability to capture accurate event probabilities. However, its evaluation can be affected by the integration range and the time density of available samples.

The third most common metric for survival models is the Cumulative AUC. While the AUC is traditionally a classification metric, its application extends to survival studies with time-dependent outcomes [16]. In this context, the AUC examines the predicted survival probabilities against observed event statuses over several time instants. Samples that are censored before or during this period are treated as negative events. The Cumulative AUC integrates these time-dependent AUC values, with 1 indicating perfection in prediction.

To adjust for censoring biases, the Inverse Probability of Censoring Weighting (IPCW) method [14, 17] is employed. Here, each sample is assigned a weight based on its inverse censoring probability at a given time. Observations with high censoring likelihoods get more weight, and vice versa for low-censoring observations. This weighting helps to counteract potential biases due to the event censoring distribution. Also, each of the metrics described focuses on a specific aspect of survival models. Therefore, for a comprehensive evaluation of the overall quality of a survival model, multiple metrics must be taken into account.

3. Related Work

In recent years, deep learning increased the expressive capability of traditional survival models. The first works were devoted to the extension of one of the most prominent survival models: the Cox model [12]. The Cox model defines a hazard function based on the assumption that the relative risk between subjects remains unchanged over time (proportional hazard assumption):

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i^T\beta), \tag{4}$$

where $h_0(t)$ is the baseline hazard common across all subjects, and $\exp(\mathbf{x}_i^T\beta)$ is a subject-specific factor that modifies the baseline hazard based on an individual's risk profile. The classic Cox model assumes the existence of a linear relationship between features and subject hazard with the risk multiplier being the exponential of the dot product of features and weights.

A substantial extension of the Cox model is DeepSurv [2]. Here, the linear relationship between features and risks is replaced with a deep neural network, capturing non-linear interactions between features and the hazard function. It leverages the same differentiable loss function as the original Cox model for training, called partial log-likelihood. This loss function is tailored to train models based on the proportional hazard assumption.

However, the proportional hazards assumption, though rendering models straightforward and interpretable, can sometimes hamper their generalization. In fact, many real-world datasets do not respect this assumption, rendering such models less effective. A paradigm shift in neural survival models emerged with time discretization techniques [5]. These techniques allowed neural networks to directly approximate discretized hazard and survival functions. Among the models following this approach, DeepHit [3] employs sigmoid activations to estimate discrete probabilities for designated event times. DeepHit is specifically tailored to compute probabilities for multiple competing events, predicting which event occurs first. In fact, its loss function is designed not only to improve the model's accuracy but also to predict event occurrence in the most probable order.

Drawing inspiration from the Multi-Task Logistic Regression (MTLR) approach [18], Neural Multi-Task Logistic Regression (N-MTLR) [19] employs multiple neural-based logistic regression heads to predict event occurrence probability for each time step. These outputs are subsequently normalized using a softmax function to yield event probabilities.

Finally, the Logistic Hazard model [20, 5] frames the survival problem discretely, transforming it into a sequence of binary classification tasks. Each task predicts the risk for an event occurrence at a given time interval. The model captures time-dependent effects through a multi-output neural network employing softmax activations, making it a robust choice for handling time-varying effects in survival analysis.

An alternative approach from [5, 10] instead of discretizing the survival function, assumes the hazard function to be piece-wise constant. This method, called PC-Hazard, produces continuous survival functions framed as piece-wise exponentials. Thus, PC-Hazard adapts any regression model to a survival model, trainable with the Poisson regression technique.

4. Interpolation Methods

In survival models based on neural networks, the discrete outputs, or *anchor points*, define the value of the survival function for a set of specific time instants. This section provides a description of several interpolation techniques designed to bridge the gap between discrete survival functions and continuous metric evaluation. Consider a set of B time instants, each corresponding to the limit of a discretization bin, $\{\tau_1, \tau_2, \dots, \tau_B\}$, such that $0 < \tau_1 < \tau_2 < \cdots < \tau_B$. Then, a survival model based on neural networks produces a set of outputs $\{s_1, s_2, \dots, s_B\}$, such that $1 \geq s_1 \geq s_2 \geq \cdots \geq s_B \geq 0$. The set of pairs (τ_i, s_i) corresponds to the anchor points leveraged by the interpolation methods to obtain a continuous survival function. In order to allow the interpolation to attain the properties of survival functions, we consider the pairs $(0, 1)$ and $(\tau_\infty, 0)$ to always be part of the set of anchor points. Figure 1 illustrates the considered interpolation techniques evaluated on a set of fixed anchor points.

4.1. Step-wise Interpolation

Most works apply step-wise interpolation to produce continuous outputs from the set of anchor points produced by a survival model. In particular, given a time instant $t \in [\tau_i, \tau_{i+1})$, this simple

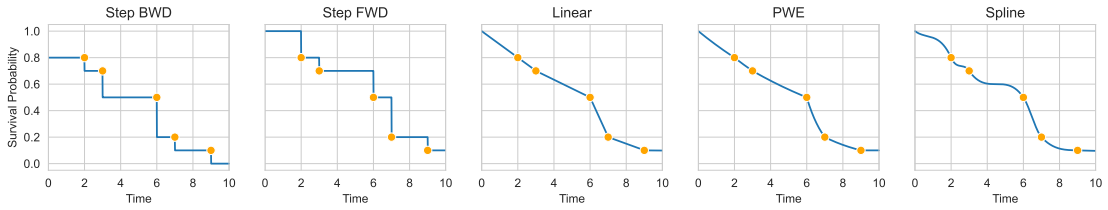


Figure 1: Interpolation techniques involved in this study. The fixed anchor points are drawn in yellow for $t = 2, 3, 6, 7$, and 9. The first two plots from the left refer to step-wise interpolations considering the following or previous anchor points, respectively. The third plot illustrates a linear interpolation. The fourth is a piece-wise exponential, inspired by the PC-Hazard model [5, 10]. The final plot interpolates the anchor points with a monotonic cubic spline.

type of interpolation defines the value of a survival function as

$$S(t) = s_i \tag{5}$$

which corresponds to the value of the closest anchor points with a lower corresponding time. We call this interpolation method *Step FWD*, indicating that the anchor point is propagated forward in the survival function. In the following analyses, we also employ an alternative approach, called *Step BWD*, which propagates the next closest anchor point backward in the survival function as

$$S(t) = s_{i+1}. \tag{6}$$

The idea of *Step BWD* is to focus on future event instances rather than the immediate past. It might be relevant in situations where interventions or treatments are planned, and the anticipation of the next event risk is more clinically significant than the immediate past.

4.2. Linear Interpolation

The most straightforward extension to the step-wise interpolation techniques is to define the interpolation point on the line connecting the considered anchor points. A linearly-interpolated survival function is defined as

$$S(t) = s_i + \frac{t - t_i}{t_{i+1} - t_i} (s_{i+1} - s_i). \tag{7}$$

4.3. Piece-wise Exponential Interpolation

This interpolation method is inspired by the piece-wise constant hazard model [5, 10]. This method, referred to as *PWE*, assumes the hazard function to be constant within each time interval. Then, according to Eq. (3), the survival function results in a piece-wise exponential function. The interpolation is computed as

$$S(t) = s_i \exp\left(\lambda_i \cdot \frac{t - t_i}{t_{i+1} - t_i}\right) \tag{8}$$

where

$$\lambda_i = \ln\left(\frac{s_{i+1}}{s_i}\right). \tag{9}$$

4.4. Monotonic Cubic Spline Interpolation

The Hermite spline with monotonicity constraints [21] is a spline-based interpolation method to fit a set of anchor points with a non-increasing smooth function maintaining a continuous derivative. The Fritsch–Carlson method enables the construction of survival functions with a smooth transition between anchor points. In the subsequent sections, we refer to this interpolation technique as *Spline*.

The idea is to constrain the tangents of the Hermit spline in such a way that the resulting piece-wise function is monotonic. To this end, the Fritsch–Carlson method starts from the secant lines between successive anchor points

$$\delta_i = \frac{s_{i+1} - s_i}{t_{i+1} - t_i} \tag{10}$$

and initializes the average of the secants as

$$m_i = \frac{1}{2}(\delta_{i-1} + \delta_i), \tag{11}$$

assuming $m_1 = \delta_1$ and $m_B = \delta_B$. For pairs of anchors where $s_i = s_{i+1}$, let $m_i = 0$. For all the other pairs, instead, let

$$\alpha_i = \frac{m_i}{\delta_i} \quad \text{and} \quad \beta_i = \frac{m_{i+1}}{\delta_i}. \tag{12}$$

A sufficient condition to ensure monotonicity is to set, for $\alpha_i > 3$ or $\beta_i > 3$,

$$m_i = 3\delta_i. \tag{13}$$

At this point, the values of tangents m_i guarantee that a Hermit spline passing through the anchor points is non-increasing. The survival function is computed as

$$S(\mu) = (2\mu^3 - 3\mu^2 + 1)s_i + (\mu^3 - 2\mu^2 + \mu)m_i + (-2\mu^3 + 3\mu^2)s_{i+1} + (\mu^3 - \mu^2)m_{i+1} \tag{14}$$

where

$$\mu = \frac{t - t_i}{t_{i+1} - t_i}. \tag{15}$$

5. Experiments

This section collects the experimental methodology and results to validate our claims about interpolation techniques for neural-based survival models. In particular, we describe the datasets involved in the experiments, the training and inference procedure, and the final results obtained. We highlight that each of the datasets involved is publicly accessible and used in several survival studies for benchmarking purposes [22, 8, 9]. To allow for reproducibility, we made the source code of the experiments publicly available[1].

[1] https://github.com/archettialberto/interpolation_for_deep_survival_analysis

Table 1
Summary statistics of the survival datasets involved in the experiments.

Dataset	Samples	Censored	Numerical Features	Categorical Features
WHAS500 [23]	461	38%	7	9
GBSG2 [24]	686	44%	5	3
METABRIC [25, 2]	1904	58%	5	3
TCGA-BRCA [26]	1048	14%	1	38

5.1. Datasets

This section describes the datasets processed throughout the experiments:

- **Worcester Heart Attack Study (WHAS500)** [23]: This dataset focuses on cardiovascular health, specifically patients who have experienced myocardial infarction. Given that heart diseases are one of the leading causes of mortality worldwide, models built on this dataset can help in risk prediction, better understanding of prognostic factors, and overall improved patient management strategies.
- **German Breast Cancer Study Group (GBSG2)** [24]: Cancer recurrence is a significant concern for patients who have undergone treatment. The GBSG2 dataset provides insights into factors that may affect recurrence, especially in the context of hormone treatments. The dataset's focus on covariates like age, menopausal status, and tumor-specific details makes it a rich source for modeling and predictions, which can directly influence treatment decisions.
- **Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)** [25, 2]: This dataset offers clinical attributes related to patients experiencing breast cancer. It is part of a larger project offering genomic data, paving the way for personalized treatment plans by taking into account the genetic variations that might influence survival rates.
- **The Cancer Genome Atlas Program - Breast Cancer Study (TCGA-BRCA)** [26]: The TCGA provides a comprehensive view of the genomic changes across various cancer types. Among the data collection projects revolving around TCGA, BRCA focuses on breast-invasive carcinoma, offering insights into the variations in survival outcomes based on geographic regions and their associated clinical practices. This dataset comes from a dataset suite for medical federated learning, called Flamby [26]. In this study, we do not consider the federated aspect, aggregating the regional clients into a single cluster of individuals.

Table 1 collects the summary statistics of these datasets, with a focus on their sample dimensionality, censorship percentages, and feature types.

5.2. Experimental Setup

This section delineates the methodological approach utilized to assess the efficacy of interpolation as a post-processing measure in survival models based on neural networks. The datasets

employed for our evaluation, specifically WHAS500, GBSG2, METABRIC, and TCGA-BRCA, are detailed in Section 5.1. Data from these datasets were uniformly sampled to formulate both training and test splits, comprising 80% and 20% of the overall samples, respectively. Subsequently, the training subset underwent an additional 80-20% split to generate a validation subset.

The experiments involved four state-of-the-art neural network-based models from survival analysis: DeepSurv, DeepHit, Logistic Hazard, and N-MTLR, each thoroughly described in Section 3. Notably, DeepSurv is the only model based on the proportional hazard assumption, whereas the others rely on an explicit definition of discrete time bins. Concerning these discretization points, we adopted a uniform splitting approach, increasing the anchor count with every experiment. The tested numbers of anchors are 5, 10, 50, 100, 500, and 1000. These numbers hold for non-proportional models only, as DeepSurv has a fixed number of anchors, corresponding to the points of the baseline function, shared across all subjects.

Each model comprises a two-layer fully connected neural network with a number of inputs equal to the dataset features and a hidden layer size of 32. Each layer is followed by a ReLU activation function and a dropout regularization layer with 0.1 probability. The number of outputs is 1 for DeepSurv and equal to the number of anchor points for all the other models. In the experiments, models are trained using the Adam optimizer with a learning rate of 0.01. Training executed till convergence for a maximum of 300 epochs, adopting an early stopping strategy on the validation set with a 10-epoch patience threshold. The selected batch size was fixed at 128.

In the subsequent inference phase, survival functions were derived from the anchor points of each model, after an interpolation step leveraging the methods outlined in Section 4 – Step BWD, Step FWD, Linear, PWE, and Spline. For each trained model paired with an interpolation strategy, the C-Index, the IBS, and the Cumulative AUC with IPCW weighting were evaluated, as described in Section 2.2. The IBS and the Cumulative AUC were integrated over the 25th and 75th percentiles of the test times, to limit the noise that could be introduced by the lower sample density at the endpoints of the time spectrum. Finally, to limit the effects of randomness, each single experiment was repeated 30 times, averaging the final results.

5.3. Results

In this section, we present and discuss the empirical results derived from our experiments with various interpolation techniques. For brevity, we enumerate the IBS (Table 2a), Cumulative AUC (Table 2b), and C-Index (Table 2c) values achieved on the METABRIC dataset, which is the largest dataset among the ones analyzed, for 10, 100, and 1000 anchor points. Detailed numerical values on the WHAS500, GBSG2, and TCGA-BRCA datasets are reported in Appendix A. On top of that, the time-dependent metrics for all datasets, namely IBS and Cumulative AUC, are plotted for 5, 50, and 500 anchor counts in Figure 2a and Figure 2b.

Does interpolation serve as an effective post-processing step when evaluated using the IBS metric? As illustrated in Table 2a and Figure 2a, implementing any form of interpolation generally proves beneficial over the Step BWD or Step FWD techniques. Specifically, for a limited number of anchor points, i.e., 5 and 10, neural models leveraging Linear and PWE interpolations demon-

strate a better IBS compared to their counterparts. Although Spline interpolation surpasses step-wise methods, it falls behind Linear and PWE. As the number of anchor points increases, the distinction among interpolation methods diminishes. This is expected, as a larger anchor count offers a finer discretization grid, enabling the neural network to precisely adjust the survival function and thereby mitigating the necessity for interpolation. Notably, while minor differences can still be observed at 50 and 100 anchors, increasing to 500 or 1000 effectively equalizes the results of all methods. This convergence can be attributed to the anchor count approaching the dataset size, compelling the model to capture the behavior of individual time instances.

Does interpolation serve as an effective post-processing step when evaluated using the Cumulative AUC metric? The Cumulative AUC metric outcomes, reported in Table 2b and Figure 2b, largely follow the trend of the previous observations. Non-step-based interpolation methods tend to augment the Cumulative AUC for neural models, especially when the number of anchor points is low. An outlier to this trend is observed with DeepHit using 10 anchor points on the METABRIC dataset, where Step FWD emerges as the best technique. However, this remains the only exception with respect to the general trend. Remarkably, while Step FWD often serves as a default choice for state-of-the-art survival models, it is consistently outperformed by Step BWD. Similar to the IBS trend, the performance difference among interpolation techniques diminishes with an increased anchor count.

How do interpolation techniques affect the C-Index metric? As highlighted in Table 2c, step-based interpolation methods marginally outperform other techniques regarding the C-Index on the METABRIC dataset. Hence, for specific applications where concordance is the only metric of utmost importance, step-based interpolation stands as a reliable choice. On the other hand, for any other situation, smoother interpolation techniques present better time-dependent metrics with only a negligible degradation of concordance.

Is there a correlation between the proportional hazard assumption and interpolation's efficacy? The proportional hazard assumption significantly impacts the model's outputs, imposing constant subject ratios over time. Consequently, the chosen interpolation method should not affect the C-Index, as confirmed by DeepSurv's performance in Table 2c. Interestingly, for the other metrics, IBS and Cumulative AUC, deviations are not noticeable to the fourth decimal place. Thus, for models based on the proportional hazard assumption, the influence of interpolation on performance is negligible. Instead, as thoroughly analyzed earlier, the opposite holds for non-proportional models based on time discretization.

How does censoring impact results? As previously discussed, interpolation techniques generally enhance survival metrics. This improvement is particularly evident in the METABRIC dataset, which has the most significant proportion of censored samples among the datasets we examined. When we compare this to other datasets with fewer censored samples, the positive effect of interpolation, although still present, is less marked. While it is not definitive that there is a direct correlation between interpolation and the percentage of censorship, we can affirm that a high rate of censoring does not hinder the benefits of interpolation techniques.

6. Conclusion

In this study, we investigated the influence of interpolation techniques on the performance metrics of survival models. Due to their expressive power, these models often achieve a high degree of generalization. However, their inherent discretization limitations can compromise their precision. To address this, we focused on the post-processing of survival functions through interpolation between anchor points, aiming to improve time-dependent metrics such as the IBS and Cumulative AUC. The empirical analyses conducted across various real-world healthcare datasets and model configurations underscored a consistent pattern: even simple interpolation methods, like linear interpolation, offer tangible improvements in these metrics. This trend is especially noticeable when the number of anchor points is orders of magnitude smaller than the dataset cardinality, which corresponds to most real-world use cases. In summary, this study underscores the potential of combining the expressiveness of neural networks with interpolation techniques to improve the accuracy of survival predictions in clinical contexts.

7. Ethical Discussion

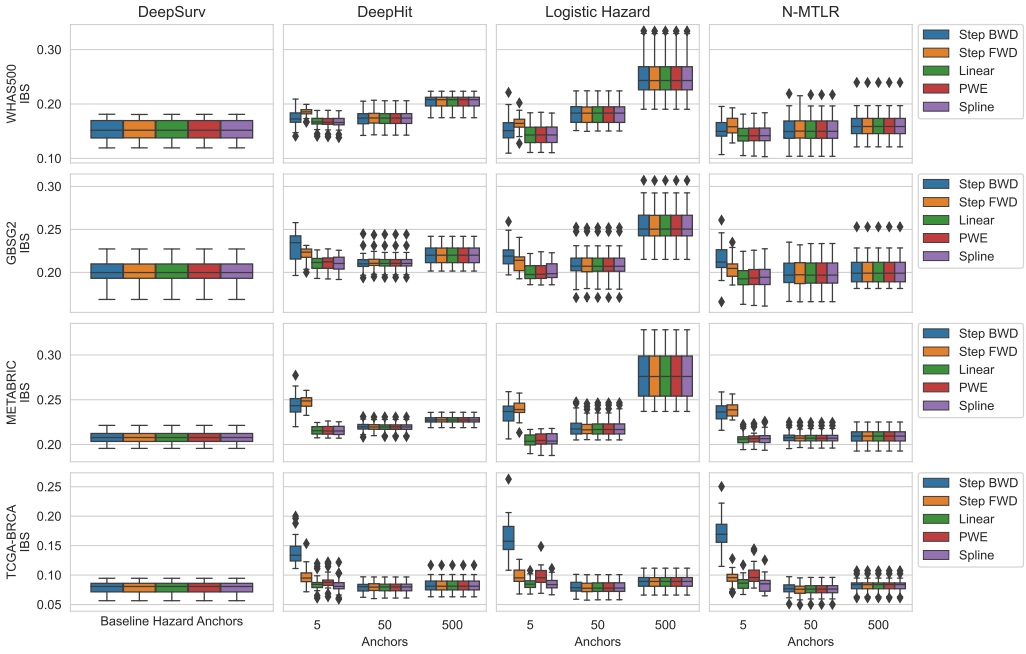
While our study focuses on a specific mathematical question concerning the post-processing of existing, well-studied survival models, the delicate nature of risk assessment in the healthcare domain raises discussions on several ethical dimensions. First, at its core, survival analysis studies the probability outcomes of events over time. In the medical field, the results of SA models may influence decision-making and treatment priorities. The potential prioritization of patients based solely on statistical outcomes may lead to short-sighted decisions. Therefore, the outcomes of survival models should be used as suggestions for domain experts who must take actions based on several real-world factors that may inevitably not be captured by statistical models.

Second, the use of patient data must undergo consent and transparency. Especially in the healthcare domain, where data are sensitive and privacy-protected, it is of utmost importance to ensure that the rights of individuals and data owners are respected. In this study, we utilized publicly available survival datasets that are commonly used to benchmark survival techniques.

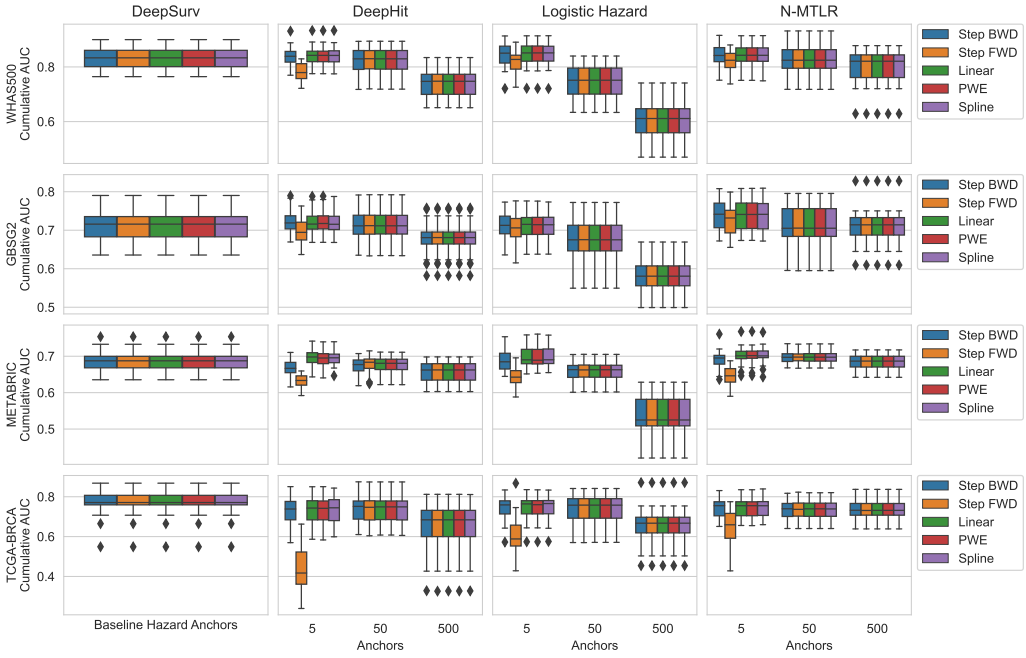
In conclusion, while our focus specifically addresses a technical aspect of survival models, we recognize the broader impact of survival analysis. Our hope is that by enhancing the reliability of these models, we contribute to a more ethical and fair healthcare landscape where statistical predictions serve as one tool among many, to aid judgments of medical professionals.

Acknowledgments

This project has been supported by AI-SPRINT: AI in Secure Privacy-pReserving computINg conTinuum (European Union H2020 grant agreement No. 101016577) and FAIR: Future Artificial Intelligence Research (NextGenerationEU, PNRR-PE-AI scheme, M4C2, investment 1.3, line on Artificial Intelligence).



(a) IBS results.



(b) Cumulative AUC results.

Figure 2: IBS and Cumulative AUC on all datasets for 5, 50, and 500 anchor points. For IBS, the lower the better; for the Cumulative AUC, the higher the better. Columns correspond to survival models, while rows correspond to survival datasets. Results are averaged over 30 runs.

Table 2
Results based on IBS (Table 2a), Cumulative AUC (Table 2b), and C-Index (Table 2c) considering the METABRIC dataset for 10, 100, and 1000 anchor points. Results are averaged over 30 runs and scaled up by a factor of 100 for better readability.

(a) IBS mean and standard deviation.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7
DeepHit	10	22.1 ± 0.6	22.2 ± 0.4	**21.6 ± 0.4**	**21.6 ± 0.4**	**21.6 ± 0.4**
Logistic Hazard	10	21.7 ± 0.8	21.5 ± 0.7	**21.0 ± 0.7**	**21.0 ± 0.7**	21.1 ± 0.7
N-MTLR	10	21.3 ± 1.0	21.1 ± 0.7	**20.6 ± 0.9**	**20.6 ± 0.9**	20.7 ± 0.9
DeepHit	100	22.1 ± 0.6	**22.0 ± 0.6**	**22.0 ± 0.6**	**22.0 ± 0.6**	**22.0 ± 0.6**
Logistic Hazard	100	22.9 ± 1.1	**22.8 ± 1.1**	**22.8 ± 1.1**	**22.8 ± 1.1**	**22.8 ± 1.1**
N-MTLR	100	21.1 ± 0.9	21.1 ± 0.8	21.1 ± 0.8	21.1 ± 0.8	21.1 ± 0.8
DeepHit	1000	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5
Logistic Hazard	1000	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9
N-MTLR	1000	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6

(b) Cumulative AUC mean and standard deviation.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7
DeepHit	10	67.5 ± 2.0	**69.8 ± 1.9**	69.2 ± 1.9	69.1 ± 1.9	69.0 ± 1.9
Logistic Hazard	10	68.1 ± 2.1	68.3 ± 2.0	**68.4 ± 2.0**	**68.4 ± 2.0**	**68.4 ± 2.0**
N-MTLR	10	69.7 ± 2.3	69.9 ± 2.2	**70.0 ± 2.3**	**70.0 ± 2.3**	**70.0 ± 2.2**
DeepHit	100	66.1 ± 2.2	**66.4 ± 2.2**	66.3 ± 2.2	66.3 ± 2.2	66.3 ± 2.2
Logistic Hazard	100	63.2 ± 2.9	**63.3 ± 2.9**	**63.3 ± 2.9**	**63.3 ± 2.9**	**63.3 ± 2.9**
N-MTLR	100	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3
DeepHit	1000	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8
Logistic Hazard	1000	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6
N-MTLR	1000	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2

(c) C-Index mean and standard deviation.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0
DeepHit	10	63.4 ± 2.4	**64.3 ± 2.5**	64.0 ± 2.1	64.0 ± 2.1	64.0 ± 2.2
Logistic Hazard	10	**63.8 ± 2.0**	63.5 ± 2.0	63.7 ± 2.0	63.7 ± 2.0	63.7 ± 2.0
N-MTLR	10	**64.5 ± 1.6**	64.1 ± 1.8	64.4 ± 1.6	64.4 ± 1.6	64.3 ± 1.7
DeepHit	100	63.2 ± 2.0	63.2 ± 1.9	63.2 ± 2.0	63.2 ± 2.0	63.2 ± 2.0
Logistic Hazard	100	61.0 ± 2.5	61.0 ± 2.4	61.0 ± 2.5	61.0 ± 2.5	61.0 ± 2.5
N-MTLR	100	**63.0 ± 2.3**	62.9 ± 2.3	62.9 ± 2.3	62.9 ± 2.3	62.9 ± 2.3
DeepHit	1000	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0
Logistic Hazard	1000	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7
N-MTLR	1000	**63.4 ± 2.1**	63.3 ± 2.1	63.3 ± 2.1	63.3 ± 2.1	63.3 ± 2.1

References

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

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

[3] C. Lee, W. Zame, J. Yoon, M. Van Der Schaar, Deephit: A deep learning approach to survival analysis with competing risks, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[4] H. Kvamme, Ø. Borgan, I. Scheel, Time-to-event prediction with neural networks and cox regression, arXiv preprint arXiv:1907.00825 (2019).

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

[6] S. Wiegrebe, P. Kopper, R. Sonabend, A. Bender, Deep learning for survival analysis: A review, arXiv preprint arXiv:2305.14961 (2023).

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

[8] A. Archetti, M. Matteucci, Federated Survival Forests, in: 2023 International Joint Conference on Neural Networks (IJCNN2023), IEEE (in press), 2023.

[9] A. Archetti, F. Ieva, M. Matteucci, Scaling survival analysis in healthcare with federated survival forests: A comparative study on heart failure and breast cancer genomics, Future Generation Computer Systems 149 (2023) 343–358. URL: https://www.sciencedirect.com/science/article/pii/S0167739X23002935. doi:https://doi.org/10.1016/j.future.2023.07.036.

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

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

[12] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society. Series B (Methodological) 34 (1972) 187–220. URL: http://www.jstor.org/stable/2985181.

[13] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, M. J. Van Der Laan, Survival ensembles, Biostatistics 7 (2005) 355–373. URL: https://doi.org/10.1093/biostatistics/kxj011. doi:10.1093/biostatistics/kxj011. arXiv:https://academic.oup.com/biostatistics/article-pdf/7/3/355/690076/kxj011.pdf.

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

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

[16] S. Pölsterl, scikit-survival: A library for time-to-event analysis built on top of scikit-learn,

Journal of Machine Learning Research 21 (2020) 1–6. URL: http://jmlr.org/papers/v21/20-729.html.

[17] J. M. Robins, A. Rotnitzky, Recovery of information and adjustment for dependent censoring using surrogate markers, in: AIDS epidemiology, Springer, 1992, pp. 297–331.

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

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

[20] M. F. Gensheimer, B. Narasimhan, A scalable discrete-time survival model for neural networks, PeerJ 7 (2019) e6257.

[21] F. N. Fritsch, R. E. Carlson, Monotone piecewise cubic interpolation, SIAM Journal on Numerical Analysis 17 (1980) 238–246. URL: http://www.jstor.org/stable/2156610.

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

[23] D. W. Hosmer, S. Lemeshow, S. May, Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008. URL: http://doi.wiley.com/10.1002/9780470258019. doi:10.1002/9780470258019.

[24] M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. Neumann, H. Rauschecker, 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 (1994) 2086–2093.

[25] B. Pereira, S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut, D. W. Y. Tsui, B. Liu, S.-J. Dawson, J. Abraham, H. Northen, J. F. Peden, A. Mukherjee, G. Turashvili, A. R. Green, S. McKinney, A. Oloumi, S. Shah, N. Rosenfeld, L. Murphy, D. R. Bentley, I. O. Ellis, A. Purushotham, S. E. Pinder, A.-L. Børresen-Dale, H. M. Earl, P. D. Pharoah, M. T. Ross, S. Aparicio, C. Caldas, The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes, Nature Communications 7 (2016) 11479. URL: https://www.nature.com/articles/ncomms11479. doi:10.1038/ncomms11479.

[26] J. Ogier du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, B. Muzellec, C. Philippenko, S. Silva, M. Teleńczuk, S. Albarqouni, S. Avestimehr, A. Bellet, A. Dieuleveut, M. Jaggi, S. P. Karimireddy, M. Lorenzi, G. Neglia, M. Tommasi, M. Andreux, Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 5315–5334. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/232eee8ef411a0a316efa298d7be3c2b-Paper-Datasets_and_Benchmarks.pdf.

A. Detailed Numerical Results

This section presents all the numerical results obtained throughout our experiments. For each dataset, we list three tables, each corresponding to the IBS, Cumulative AUC, and C-Index metrics, respectively. In particular, the table-dataset correspondence is as follows:

- **WHAS500 dataset**: Table 3 (IBS), Table 4 (Cumulative AUC), and Table 5 (C-Index).
- **GBSG2 dataset**: Table 6 (IBS), Table 7 (Cumulative AUC), and Table 8 (C-Index).
- **METABRIC dataset**: Table 9 (IBS), Table 10 (Cumulative AUC), and Table 11 (C-Index).
- **TCGA-BRCA dataset**: Table 12 (IBS), Table 13 (Cumulative AUC), and Table 14 (C-Index).

A.1. WHAS500 dataset

Table 3
IBS results on the WHAS500 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	15.4 ± 2.0	15.4 ± 2.0	15.4 ± 2.0	15.4 ± 2.0	15.4 ± 2.0
DeepHit	5	17.4 ± 1.5	18.5 ± 0.8	16.8 ± 1.1	16.7 ± 1.1	16.7 ± 1.1
Logistic Hazard	5	15.4 ± 2.7	16.5 ± 1.5	14.5 ± 2.0	14.6 ± 2.1	14.5 ± 2.1
N-MTLR	5	15.2 ± 2.1	16.1 ± 1.6	14.3 ± 1.9	14.4 ± 2.0	14.4 ± 2.0
DeepHit	10	17.3 ± 1.4	17.7 ± 1.3	17.4 ± 1.4	17.4 ± 1.4	17.4 ± 1.4
Logistic Hazard	10	17.0 ± 1.9	16.9 ± 1.7	16.8 ± 1.8	16.8 ± 1.8	16.8 ± 1.8
N-MTLR	10	15.7 ± 2.1	15.5 ± 2.0	15.5 ± 2.1	15.5 ± 2.1	15.5 ± 2.1
DeepHit	50	17.4 ± 1.5	17.5 ± 1.6	17.4 ± 1.5	17.4 ± 1.5	17.4 ± 1.5
Logistic Hazard	50	18.2 ± 1.8	18.2 ± 1.8	18.2 ± 1.8	18.2 ± 1.8	18.2 ± 1.8
N-MTLR	50	15.4 ± 2.4	15.4 ± 2.4	15.4 ± 2.4	15.4 ± 2.4	15.4 ± 2.4
DeepHit	100	17.9 ± 1.3	18.0 ± 1.3	18.0 ± 1.3	18.0 ± 1.3	18.0 ± 1.3
Logistic Hazard	100	19.9 ± 2.9	19.9 ± 2.8	19.9 ± 2.9	19.9 ± 2.9	19.9 ± 2.9
N-MTLR	100	16.3 ± 2.1	16.3 ± 2.1	16.3 ± 2.1	16.3 ± 2.1	16.3 ± 2.1
DeepHit	500	20.4 ± 1.3	20.4 ± 1.3	20.4 ± 1.3	20.4 ± 1.3	20.4 ± 1.3
Logistic Hazard	500	25.0 ± 3.6	24.9 ± 3.6	25.0 ± 3.6	25.0 ± 3.6	25.0 ± 3.6
N-MTLR	500	16.2 ± 2.3	16.2 ± 2.3	16.2 ± 2.3	16.2 ± 2.3	16.2 ± 2.3
DeepHit	1000	21.6 ± 1.4	21.6 ± 1.4	21.6 ± 1.4	21.6 ± 1.4	21.6 ± 1.4
Logistic Hazard	1000	27.4 ± 3.2	27.4 ± 3.2	27.4 ± 3.2	27.4 ± 3.2	27.4 ± 3.2
N-MTLR	1000	16.5 ± 2.4	16.5 ± 2.4	16.5 ± 2.4	16.5 ± 2.4	16.5 ± 2.4

Table 4
Cumulative AUC results on the WHAS500 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	83.0 ± 4.7	83.0 ± 4.7	83.0 ± 4.7	83.0 ± 4.7	83.0 ± 4.7
DeepHit	5	83.8 ± 3.6	78.1 ± 3.2	84.0 ± 3.5	84.0 ± 3.5	84.0 ± 3.5
Logistic Hazard	5	84.2 ± 4.2	82.3 ± 4.0	84.4 ± 4.2	84.4 ± 4.2	84.4 ± 4.2
N-MTLR	5	84.4 ± 3.9	82.1 ± 3.5	84.4 ± 3.8	84.4 ± 3.8	84.4 ± 3.8
DeepHit	10	82.5 ± 3.7	82.6 ± 3.7	82.6 ± 3.7	82.6 ± 3.7	82.6 ± 3.7
Logistic Hazard	10	79.1 ± 4.3	79.2 ± 4.3	79.1 ± 4.3	79.1 ± 4.3	79.1 ± 4.3
N-MTLR	10	82.5 ± 4.6	82.4 ± 4.6	82.5 ± 4.6	82.5 ± 4.6	82.5 ± 4.6
DeepHit	50	82.3 ± 4.3	82.4 ± 4.3	82.4 ± 4.3	82.4 ± 4.3	82.4 ± 4.3
Logistic Hazard	50	74.8 ± 5.6	74.8 ± 5.6	74.8 ± 5.6	74.8 ± 5.6	74.8 ± 5.6
N-MTLR	50	82.8 ± 4.8	82.8 ± 4.8	82.8 ± 4.8	82.8 ± 4.8	82.8 ± 4.8
DeepHit	100	80.8 ± 4.1	80.9 ± 4.1	80.8 ± 4.1	80.8 ± 4.1	80.8 ± 4.1
Logistic Hazard	100	70.0 ± 6.9	70.0 ± 6.9	70.0 ± 6.9	70.0 ± 6.9	70.0 ± 6.9
N-MTLR	100	81.1 ± 4.2	81.1 ± 4.2	81.1 ± 4.2	81.1 ± 4.2	81.1 ± 4.2
DeepHit	500	74.1 ± 5.0	74.1 ± 5.0	74.1 ± 5.0	74.1 ± 5.0	74.1 ± 5.0
Logistic Hazard	500	60.3 ± 7.1	60.3 ± 7.1	60.3 ± 7.1	60.3 ± 7.1	60.3 ± 7.1
N-MTLR	500	80.4 ± 5.7	80.4 ± 5.7	80.4 ± 5.7	80.4 ± 5.7	80.4 ± 5.7
DeepHit	1000	69.5 ± 7.7	69.5 ± 7.7	69.5 ± 7.7	69.5 ± 7.7	69.5 ± 7.7
Logistic Hazard	1000	55.3 ± 7.3	55.3 ± 7.3	55.3 ± 7.3	55.3 ± 7.3	55.3 ± 7.3
N-MTLR	1000	80.4 ± 5.6	80.4 ± 5.6	80.4 ± 5.6	80.4 ± 5.6	80.4 ± 5.6

Table 5
C-Index results on the WHAS500 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	75.1 ± 5.0	75.1 ± 5.0	75.1 ± 5.0	75.1 ± 5.0	75.1 ± 5.0
DeepHit	5	74.8 ± 7.2	75.6 ± 7.2	75.5 ± 7.2	75.5 ± 7.2	75.6 ± 7.2
Logistic Hazard	5	76.0 ± 6.2	76.4 ± 6.0	76.3 ± 6.0	76.3 ± 6.0	76.3 ± 6.1
N-MTLR	5	76.4 ± 6.7	76.5 ± 6.6	76.5 ± 6.6	76.5 ± 6.6	76.5 ± 6.6
DeepHit	10	74.0 ± 5.1	74.3 ± 5.1	74.2 ± 5.1	74.2 ± 5.1	74.2 ± 5.1
Logistic Hazard	10	70.6 ± 13.0	70.7 ± 13.1	70.6 ± 13.1	70.6 ± 13.1	70.6 ± 13.1
N-MTLR	10	74.0 ± 12.7	73.9 ± 12.7	73.9 ± 12.7	73.9 ± 12.7	73.9 ± 12.7
DeepHit	50	73.4 ± 13.1	73.5 ± 13.1	73.5 ± 13.1	73.5 ± 13.1	73.5 ± 13.1
Logistic Hazard	50	67.1 ± 11.6	67.0 ± 11.6	67.0 ± 11.6	67.0 ± 11.6	67.0 ± 11.6
N-MTLR	50	72.0 ± 12.0	72.0 ± 12.1	72.0 ± 12.0	72.0 ± 12.0	72.0 ± 12.0
DeepHit	100	75.0 ± 4.7	75.1 ± 4.7	75.0 ± 4.7	75.0 ± 4.7	75.0 ± 4.7
Logistic Hazard	100	65.1 ± 12.3	65.1 ± 12.3	65.1 ± 12.3	65.1 ± 12.3	65.1 ± 12.3
N-MTLR	100	75.9 ± 6.7	75.8 ± 6.7	75.8 ± 6.7	75.8 ± 6.7	75.8 ± 6.7
DeepHit	500	67.8 ± 13.8	67.9 ± 13.8	67.8 ± 13.8	67.8 ± 13.8	67.8 ± 13.8
Logistic Hazard	500	57.5 ± 5.8	57.5 ± 5.8	57.5 ± 5.8	57.5 ± 5.8	57.5 ± 5.8
N-MTLR	500	75.2 ± 4.2	75.2 ± 4.2	75.2 ± 4.2	75.2 ± 4.2	75.2 ± 4.2
DeepHit	1000	65.9 ± 14.6	66.0 ± 14.6	66.0 ± 14.6	66.0 ± 14.6	66.0 ± 14.6
Logistic Hazard	1000	52.9 ± 5.1	52.9 ± 5.1	52.9 ± 5.1	52.9 ± 5.1	52.9 ± 5.1
N-MTLR	1000	72.9 ± 6.3	72.9 ± 6.3	72.9 ± 6.3	72.9 ± 6.3	72.9 ± 6.3

A.2. GBSG2 dataset

Table 6
IBS results on the GBSG2 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	19.4 ± 1.1	19.4 ± 1.1	19.4 ± 1.1	19.4 ± 1.1	19.4 ± 1.1
DeepHit	5	22.9 ± 1.6	22.1 ± 0.8	21.0 ± 0.9	21.0 ± 0.9	21.0 ± 0.9
Logistic Hazard	5	22.0 ± 1.4	21.2 ± 1.2	20.1 ± 1.0	20.1 ± 1.1	20.2 ± 1.1
N-MTLR	5	21.5 ± 1.9	20.5 ± 1.3	19.4 ± 1.3	19.5 ± 1.4	19.5 ± 1.4
DeepHit	10	21.3 ± 1.2	21.2 ± 1.1	21.0 ± 1.1	21.0 ± 1.1	21.0 ± 1.1
Logistic Hazard	10	20.4 ± 1.6	20.2 ± 1.2	20.1 ± 1.4	20.1 ± 1.4	20.1 ± 1.4
N-MTLR	10	19.9 ± 1.5	19.8 ± 1.1	19.6 ± 1.3	19.6 ± 1.3	19.7 ± 1.3
DeepHit	50	21.2 ± 1.0	21.2 ± 0.9	21.2 ± 1.0	21.2 ± 1.0	21.2 ± 1.0
Logistic Hazard	50	20.9 ± 1.8	20.9 ± 1.7	20.9 ± 1.7	20.9 ± 1.7	20.9 ± 1.7
N-MTLR	50	19.9 ± 1.6	19.9 ± 1.6	19.9 ± 1.6	19.9 ± 1.6	19.9 ± 1.6
DeepHit	100	21.6 ± 1.3	21.6 ± 1.3	21.6 ± 1.3	21.6 ± 1.3	21.6 ± 1.3
Logistic Hazard	100	21.6 ± 1.9	21.6 ± 1.9	21.6 ± 1.9	21.6 ± 1.9	21.6 ± 1.9
N-MTLR	100	19.8 ± 1.3	19.8 ± 1.3	19.8 ± 1.3	19.8 ± 1.3	19.8 ± 1.3
DeepHit	500	22.0 ± 1.1	22.0 ± 1.1	22.0 ± 1.1	22.0 ± 1.1	22.0 ± 1.1
Logistic Hazard	500	25.5 ± 2.2	25.4 ± 2.2	25.5 ± 2.2	25.5 ± 2.2	25.5 ± 2.2
N-MTLR	500	20.3 ± 1.6	20.3 ± 1.6	20.3 ± 1.6	20.3 ± 1.6	20.3 ± 1.6
DeepHit	1000	22.2 ± 1.0	22.2 ± 1.0	22.2 ± 1.0	22.2 ± 1.0	22.2 ± 1.0
Logistic Hazard	1000	26.2 ± 2.5	26.2 ± 2.5	26.2 ± 2.5	26.2 ± 2.5	26.2 ± 2.5
N-MTLR	1000	20.3 ± 1.4	20.3 ± 1.4	20.3 ± 1.4	20.3 ± 1.4	20.3 ± 1.4

Table 7
Cumulative AUC results on the GBSG2 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	73.2 ± 3.4	73.2 ± 3.4	73.2 ± 3.4	73.2 ± 3.4	73.2 ± 3.4
DeepHit	5	72.1 ± 2.8	69.8 ± 3.5	72.3 ± 2.9	72.3 ± 2.9	72.2 ± 2.9
Logistic Hazard	5	71.0 ± 3.6	70.1 ± 3.8	71.0 ± 3.7	71.0 ± 3.7	70.9 ± 3.7
N-MTLR	5	73.8 ± 3.9	72.7 ± 3.9	73.7 ± 3.9	73.7 ± 3.9	73.7 ± 3.9
DeepHit	10	70.6 ± 4.5	70.7 ± 4.3	70.8 ± 4.4	70.8 ± 4.4	70.7 ± 4.3
Logistic Hazard	10	70.5 ± 4.2	70.6 ± 4.2	70.5 ± 4.2	70.5 ± 4.2	70.5 ± 4.2
N-MTLR	10	72.1 ± 3.9	71.9 ± 3.8	72.0 ± 3.8	72.0 ± 3.8	72.0 ± 3.8
DeepHit	50	71.2 ± 3.7	71.2 ± 3.7	71.2 ± 3.7	71.2 ± 3.7	71.2 ± 3.7
Logistic Hazard	50	67.4 ± 5.3	67.4 ± 5.3	67.4 ± 5.3	67.4 ± 5.3	67.4 ± 5.3
N-MTLR	50	71.3 ± 4.4	71.3 ± 4.4	71.3 ± 4.4	71.3 ± 4.4	71.3 ± 4.4
DeepHit	100	70.0 ± 4.6	70.0 ± 4.6	70.0 ± 4.6	70.0 ± 4.6	70.0 ± 4.6
Logistic Hazard	100	66.2 ± 5.2	66.2 ± 5.2	66.2 ± 5.2	66.2 ± 5.2	66.2 ± 5.2
N-MTLR	100	71.7 ± 4.5	71.7 ± 4.5	71.7 ± 4.5	71.7 ± 4.5	71.7 ± 4.5
DeepHit	500	68.0 ± 4.0	68.0 ± 4.0	68.0 ± 4.0	68.0 ± 4.0	68.0 ± 4.0
Logistic Hazard	500	58.3 ± 4.4	58.3 ± 4.4	58.3 ± 4.4	58.3 ± 4.4	58.3 ± 4.4
N-MTLR	500	70.9 ± 4.0	70.9 ± 4.0	70.9 ± 4.0	70.9 ± 4.0	70.9 ± 4.0
DeepHit	1000	66.0 ± 4.7	66.0 ± 4.7	66.0 ± 4.7	66.0 ± 4.7	66.0 ± 4.7
Logistic Hazard	1000	58.3 ± 5.7	58.3 ± 5.7	58.3 ± 5.7	58.3 ± 5.7	58.3 ± 5.7
N-MTLR	1000	69.7 ± 4.1	69.7 ± 4.1	69.7 ± 4.1	69.7 ± 4.1	69.7 ± 4.1

Table 8
C-Index results on the GBSG2 dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	67.6 ± 5.2	67.6 ± 5.2	67.6 ± 5.2	67.6 ± 5.2	67.6 ± 5.2
DeepHit	5	66.6 ± 4.5	66.6 ± 4.5	66.9 ± 4.5	66.8 ± 4.5	67.0 ± 4.3
Logistic Hazard	5	66.5 ± 3.7	66.4 ± 3.8	66.4 ± 3.7	66.4 ± 3.7	66.4 ± 3.8
N-MTLR	5	67.5 ± 3.6	67.5 ± 3.6	67.6 ± 3.6	67.6 ± 3.6	67.6 ± 3.6
DeepHit	10	66.6 ± 3.8	66.8 ± 3.8	66.7 ± 3.8	66.7 ± 3.8	66.7 ± 3.9
Logistic Hazard	10	68.1 ± 6.2	68.1 ± 6.2	68.1 ± 6.2	68.1 ± 6.2	68.1 ± 6.2
N-MTLR	10	67.8 ± 5.4	67.6 ± 5.3	67.8 ± 5.4	67.8 ± 5.4	67.8 ± 5.4
DeepHit	50	66.7 ± 5.2	66.7 ± 5.0	66.6 ± 5.0	66.6 ± 5.0	66.6 ± 5.0
Logistic Hazard	50	64.3 ± 5.5	64.3 ± 5.5	64.3 ± 5.5	64.3 ± 5.5	64.3 ± 5.5
N-MTLR	50	65.8 ± 4.4	65.8 ± 4.5	65.8 ± 4.4	65.8 ± 4.4	65.8 ± 4.4
DeepHit	100	66.6 ± 5.7	66.6 ± 5.7	66.6 ± 5.7	66.6 ± 5.7	66.6 ± 5.7
Logistic Hazard	100	63.9 ± 5.1	63.9 ± 5.1	63.9 ± 5.1	63.9 ± 5.1	63.9 ± 5.1
N-MTLR	100	66.4 ± 5.2	66.4 ± 5.2	66.4 ± 5.2	66.4 ± 5.2	66.4 ± 5.2
DeepHit	500	63.6 ± 5.5	63.6 ± 5.5	63.6 ± 5.5	63.6 ± 5.5	63.6 ± 5.5
Logistic Hazard	500	58.2 ± 5.2	58.2 ± 5.2	58.2 ± 5.2	58.2 ± 5.2	58.2 ± 5.2
N-MTLR	500	66.0 ± 5.5	66.0 ± 5.5	66.0 ± 5.5	66.0 ± 5.5	66.0 ± 5.5
DeepHit	1000	62.4 ± 4.4	62.4 ± 4.4	62.4 ± 4.4	62.4 ± 4.4	62.4 ± 4.4
Logistic Hazard	1000	59.2 ± 7.7	59.2 ± 7.7	59.2 ± 7.7	59.2 ± 7.7	59.2 ± 7.7
N-MTLR	1000	67.0 ± 5.6	67.0 ± 5.6	67.0 ± 5.6	67.0 ± 5.6	67.0 ± 5.6

A.3. METABRIC dataset

Table 9
IBS results on the METABRIC dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7	21.0 ± 0.7
DeepHit	5	24.5 ± 1.2	24.8 ± 0.7	21.6 ± 0.5	21.6 ± 0.5	21.6 ± 0.5
Logistic Hazard	5	23.4 ± 1.3	24.0 ± 0.9	20.5 ± 0.8	20.5 ± 0.9	20.6 ± 0.9
N-MTLR	5	23.7 ± 1.1	23.9 ± 0.8	20.6 ± 0.7	20.7 ± 0.7	20.7 ± 0.7
DeepHit	10	22.1 ± 0.6	22.2 ± 0.4	21.6 ± 0.4	21.6 ± 0.4	21.6 ± 0.4
Logistic Hazard	10	21.7 ± 0.8	21.5 ± 0.7	21.0 ± 0.7	21.0 ± 0.7	21.1 ± 0.7
N-MTLR	10	21.3 ± 1.0	21.1 ± 0.7	20.6 ± 0.9	20.6 ± 0.9	20.7 ± 0.9
DeepHit	50	22.0 ± 0.5	22.0 ± 0.5	22.0 ± 0.5	22.0 ± 0.5	22.0 ± 0.5
Logistic Hazard	50	22.0 ± 1.1	21.9 ± 1.0	21.9 ± 1.1	21.9 ± 1.1	21.9 ± 1.1
N-MTLR	50	20.7 ± 0.7	20.7 ± 0.6	20.7 ± 0.7	20.7 ± 0.7	20.7 ± 0.7
DeepHit	100	22.1 ± 0.6	22.0 ± 0.6	22.0 ± 0.6	22.0 ± 0.6	22.0 ± 0.6
Logistic Hazard	100	22.9 ± 1.1	22.8 ± 1.1	22.8 ± 1.1	22.8 ± 1.1	22.8 ± 1.1
N-MTLR	100	21.1 ± 0.9	21.1 ± 0.8	21.1 ± 0.8	21.1 ± 0.8	21.1 ± 0.8
DeepHit	500	22.7 ± 0.4	22.8 ± 0.4	22.7 ± 0.4	22.7 ± 0.4	22.7 ± 0.4
Logistic Hazard	500	27.8 ± 2.8	27.8 ± 2.8	27.8 ± 2.8	27.8 ± 2.8	27.8 ± 2.8
N-MTLR	500	21.0 ± 0.8	21.0 ± 0.8	21.0 ± 0.8	21.0 ± 0.8	21.0 ± 0.8
DeepHit	1000	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5	23.1 ± 0.5
Logistic Hazard	1000	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9	30.9 ± 1.9
N-MTLR	1000	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6	21.3 ± 0.6

Table 10
Cumulative AUC results on the METABRIC dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7	68.0 ± 2.7
DeepHit	5	66.9 ± 2.2	63.1 ± 2.0	69.7 ± 2.1	69.5 ± 2.1	69.5 ± 2.0
Logistic Hazard	5	68.7 ± 3.1	64.0 ± 2.6	69.8 ± 3.0	69.8 ± 3.0	69.9 ± 2.8
N-MTLR	5	69.1 ± 2.4	64.6 ± 2.7	70.2 ± 2.4	70.2 ± 2.4	70.3 ± 2.3
DeepHit	10	67.5 ± 2.0	69.8 ± 1.9	69.2 ± 1.9	69.1 ± 1.9	69.0 ± 1.9
Logistic Hazard	10	68.1 ± 2.1	68.3 ± 2.0	68.4 ± 2.0	68.4 ± 2.0	68.4 ± 2.0
N-MTLR	10	69.7 ± 2.3	69.9 ± 2.2	70.0 ± 2.3	70.0 ± 2.3	70.0 ± 2.2
DeepHit	50	67.1 ± 2.4	67.6 ± 2.3	67.4 ± 2.4	67.4 ± 2.4	67.4 ± 2.4
Logistic Hazard	50	65.6 ± 2.5	65.7 ± 2.6	65.7 ± 2.6	65.7 ± 2.6	65.7 ± 2.6
N-MTLR	50	69.7 ± 1.6	69.7 ± 1.5	69.7 ± 1.6	69.7 ± 1.6	69.7 ± 1.6
DeepHit	100	66.1 ± 2.2	66.4 ± 2.2	66.3 ± 2.2	66.3 ± 2.2	66.3 ± 2.2
Logistic Hazard	100	63.2 ± 2.9	63.3 ± 2.9	63.3 ± 2.9	63.3 ± 2.9	63.3 ± 2.9
N-MTLR	100	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3	68.3 ± 2.3
DeepHit	500	65.6 ± 2.8	65.6 ± 2.8	65.6 ± 2.8	65.6 ± 2.8	65.6 ± 2.8
Logistic Hazard	500	53.9 ± 4.9	53.9 ± 4.9	53.9 ± 4.9	53.9 ± 4.9	53.9 ± 4.9
N-MTLR	500	68.4 ± 2.2	68.4 ± 2.2	68.4 ± 2.2	68.4 ± 2.2	68.4 ± 2.2
DeepHit	1000	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8	63.6 ± 2.8
Logistic Hazard	1000	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6	49.0 ± 3.6
N-MTLR	1000	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2	67.8 ± 2.2

Table 11
C-Index results on the METABRIC dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0	64.5 ± 2.0
DeepHit	5	63.7 ± 3.0	63.3 ± 4.0	64.7 ± 2.6	64.5 ± 2.6	64.6 ± 2.5
Logistic Hazard	5	64.2 ± 2.9	63.0 ± 3.0	64.0 ± 2.9	64.1 ± 2.9	63.9 ± 2.9
N-MTLR	5	63.6 ± 4.0	63.0 ± 4.2	63.7 ± 4.0	63.7 ± 4.0	63.6 ± 4.0
DeepHit	10	63.4 ± 2.4	64.3 ± 2.5	64.0 ± 2.1	64.0 ± 2.1	64.0 ± 2.2
Logistic Hazard	10	63.8 ± 2.0	63.5 ± 2.0	63.7 ± 2.0	63.7 ± 2.0	63.7 ± 2.0
N-MTLR	10	64.5 ± 1.6	64.1 ± 1.8	64.4 ± 1.6	64.4 ± 1.6	64.3 ± 1.7
DeepHit	50	63.5 ± 2.7	63.8 ± 2.7	63.7 ± 2.7	63.6 ± 2.7	63.6 ± 2.7
Logistic Hazard	50	62.8 ± 2.3	62.7 ± 2.4	62.7 ± 2.3	62.7 ± 2.3	62.7 ± 2.3
N-MTLR	50	64.2 ± 3.3	64.1 ± 3.2	64.1 ± 3.2	64.1 ± 3.2	64.1 ± 3.2
DeepHit	100	63.2 ± 2.0	63.2 ± 1.9	63.2 ± 2.0	63.2 ± 2.0	63.2 ± 2.0
Logistic Hazard	100	61.0 ± 2.5	61.0 ± 2.4	61.0 ± 2.5	61.0 ± 2.5	61.0 ± 2.5
N-MTLR	100	63.0 ± 2.3	62.9 ± 2.3	62.9 ± 2.3	62.9 ± 2.3	62.9 ± 2.3
DeepHit	500	61.7 ± 2.9	61.7 ± 2.9	61.7 ± 2.9	61.7 ± 2.9	61.7 ± 2.9
Logistic Hazard	500	55.2 ± 5.6	55.2 ± 5.6	55.2 ± 5.6	55.2 ± 5.6	55.2 ± 5.6
N-MTLR	500	63.1 ± 2.2	63.1 ± 2.2	63.1 ± 2.2	63.1 ± 2.2	63.1 ± 2.2
DeepHit	1000	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0	61.7 ± 3.0
Logistic Hazard	1000	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7	51.8 ± 2.7
N-MTLR	1000	63.4 ± 2.1	63.3 ± 2.1	63.3 ± 2.1	63.3 ± 2.1	63.3 ± 2.1

A.4. TCGA-BRCA dataset

Table 12
IBS results on the TCGA-BRCA dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	8.4 ± 1.1	8.4 ± 1.1	8.4 ± 1.1	8.4 ± 1.1	8.4 ± 1.1
DeepHit	5	14.1 ± 2.3	9.7 ± 1.6	8.6 ± 1.3	9.0 ± 1.4	8.3 ± 1.3
Logistic Hazard	5	16.3 ± 3.1	9.6 ± 1.6	8.5 ± 0.9	9.8 ± 1.5	8.5 ± 1.0
N-MTLR	5	17.0 ± 2.9	9.4 ± 1.3	8.6 ± 1.1	10.0 ± 1.7	8.4 ± 1.3
DeepHit	10	8.5 ± 0.7	8.6 ± 0.9	8.1 ± 0.7	8.1 ± 0.7	8.1 ± 0.7
Logistic Hazard	10	8.8 ± 1.0	8.0 ± 1.0	7.8 ± 0.9	7.8 ± 0.9	7.8 ± 0.9
N-MTLR	10	9.6 ± 1.1	8.1 ± 1.0	8.2 ± 1.0	8.2 ± 1.0	8.2 ± 1.0
DeepHit	50	7.9 ± 0.9	7.9 ± 1.0	7.9 ± 1.0	7.9 ± 1.0	7.9 ± 1.0
Logistic Hazard	50	8.0 ± 1.0	8.0 ± 1.0	8.0 ± 1.0	8.0 ± 1.0	8.0 ± 1.0
N-MTLR	50	7.6 ± 1.0	7.5 ± 1.0	7.5 ± 1.0	7.5 ± 1.0	7.5 ± 1.0
DeepHit	100	8.7 ± 1.3	8.7 ± 1.3	8.7 ± 1.3	8.7 ± 1.3	8.7 ± 1.3
Logistic Hazard	100	8.5 ± 0.9	8.5 ± 0.9	8.5 ± 0.9	8.5 ± 0.9	8.5 ± 0.9
N-MTLR	100	8.2 ± 1.3	8.2 ± 1.2	8.2 ± 1.2	8.2 ± 1.2	8.2 ± 1.3
DeepHit	500	8.3 ± 1.2	8.2 ± 1.2	8.2 ± 1.2	8.2 ± 1.2	8.2 ± 1.2
Logistic Hazard	500	9.0 ± 1.2	8.9 ± 1.2	8.9 ± 1.2	8.9 ± 1.2	8.9 ± 1.2
N-MTLR	500	8.2 ± 1.0	8.2 ± 1.0	8.2 ± 1.0	8.2 ± 1.0	8.2 ± 1.0
DeepHit	1000	8.6 ± 1.0	8.6 ± 1.0	8.6 ± 1.0	8.6 ± 1.0	8.6 ± 1.0
Logistic Hazard	1000	9.2 ± 1.3	9.2 ± 1.3	9.2 ± 1.3	9.2 ± 1.3	9.2 ± 1.3
N-MTLR	1000	8.1 ± 0.8	8.1 ± 0.8	8.1 ± 0.8	8.1 ± 0.8	8.1 ± 0.8

Table 13
Cumulative AUC results on the TCGA-BRCA dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	74.7 ± 5.7	74.7 ± 5.7	74.7 ± 5.7	74.7 ± 5.7	74.7 ± 5.7
DeepHit	5	72.9 ± 7.1	43.7 ± 10.7	73.1 ± 6.9	73.0 ± 7.0	73.3 ± 6.8
Logistic Hazard	5	74.7 ± 5.8	60.6 ± 9.8	74.8 ± 5.8	74.8 ± 5.8	74.9 ± 5.8
N-MTLR	5	74.3 ± 4.5	64.5 ± 9.0	74.4 ± 4.5	74.4 ± 4.5	74.5 ± 4.5
DeepHit	10	77.4 ± 4.4	73.3 ± 5.8	77.7 ± 4.6	77.6 ± 4.6	77.6 ± 4.8
Logistic Hazard	10	76.6 ± 6.2	76.5 ± 5.9	76.8 ± 6.1	76.8 ± 6.1	76.8 ± 6.0
N-MTLR	10	75.0 ± 6.3	74.3 ± 6.3	75.1 ± 6.2	75.0 ± 6.2	75.1 ± 6.2
DeepHit	50	74.1 ± 6.5	74.0 ± 6.6	74.1 ± 6.5	74.1 ± 6.5	74.0 ± 6.5
Logistic Hazard	50	74.0 ± 7.0	74.0 ± 7.0	74.0 ± 7.0	74.0 ± 7.0	74.0 ± 7.0
N-MTLR	50	73.4 ± 4.5	73.5 ± 4.5	73.5 ± 4.5	73.5 ± 4.5	73.5 ± 4.5
DeepHit	100	72.0 ± 6.6	72.0 ± 6.6	72.0 ± 6.6	72.0 ± 6.6	72.0 ± 6.6
Logistic Hazard	100	71.6 ± 7.6	71.6 ± 7.6	71.6 ± 7.6	71.6 ± 7.6	71.6 ± 7.6
N-MTLR	100	74.4 ± 5.7	74.4 ± 5.7	74.4 ± 5.7	74.4 ± 5.7	74.4 ± 5.7
DeepHit	500	65.9 ± 10.8	65.9 ± 10.8	65.9 ± 10.8	65.9 ± 10.8	65.9 ± 10.8
Logistic Hazard	500	65.5 ± 8.2	65.5 ± 8.2	65.5 ± 8.2	65.5 ± 8.2	65.5 ± 8.2
N-MTLR	500	73.6 ± 5.0	73.6 ± 5.0	73.6 ± 5.0	73.6 ± 5.0	73.6 ± 5.0
DeepHit	1000	63.7 ± 10.7	63.7 ± 10.8	63.7 ± 10.7	63.7 ± 10.7	63.7 ± 10.7
Logistic Hazard	1000	64.6 ± 9.8	64.6 ± 9.8	64.6 ± 9.8	64.6 ± 9.8	64.6 ± 9.8
N-MTLR	1000	72.2 ± 7.3	72.2 ± 7.3	72.2 ± 7.3	72.2 ± 7.3	72.2 ± 7.3

Table 14
C-Index results on the TCGA-BRCA dataset. Values are averaged over 30 runs and scaled up by a factor of 100 for better readability.

Model	Anchors	Step BWD	Step FWD	Linear	PWE	Spline
DeepSurv	–	68.1 ± 10.1	68.1 ± 10.1	68.1 ± 10.1	68.1 ± 10.1	68.1 ± 10.1
DeepHit	5	69.5 ± 7.5	45.3 ± 12.6	69.7 ± 7.5	69.6 ± 7.6	69.1 ± 8.1
Logistic Hazard	5	65.5 ± 11.4	56.9 ± 14.0	65.5 ± 11.4	65.5 ± 11.3	65.3 ± 11.3
N-MTLR	5	68.9 ± 12.9	64.3 ± 9.8	69.0 ± 12.9	69.0 ± 12.9	69.0 ± 12.8
DeepHit	10	70.1 ± 5.8	69.0 ± 6.7	70.0 ± 6.4	70.0 ± 6.4	69.6 ± 6.8
Logistic Hazard	10	66.5 ± 9.5	66.0 ± 9.6	66.2 ± 9.6	66.3 ± 9.6	66.3 ± 9.5
N-MTLR	10	66.5 ± 9.2	66.2 ± 9.0	66.4 ± 9.2	66.5 ± 9.2	66.4 ± 9.1
DeepHit	50	73.4 ± 10.4	73.4 ± 10.4	73.4 ± 10.4	73.4 ± 10.4	73.4 ± 10.4
Logistic Hazard	50	66.2 ± 8.1	66.2 ± 8.1	66.2 ± 8.1	66.2 ± 8.1	66.2 ± 8.1
N-MTLR	50	67.3 ± 9.6	67.2 ± 9.7	67.2 ± 9.7	67.2 ± 9.7	67.2 ± 9.7
DeepHit	100	65.6 ± 11.1	65.6 ± 11.2	65.6 ± 11.1	65.6 ± 11.1	65.6 ± 11.1
Logistic Hazard	100	62.7 ± 11.9	62.7 ± 11.9	62.7 ± 11.9	62.7 ± 11.9	62.7 ± 11.9
N-MTLR	100	70.0 ± 8.1	70.0 ± 8.1	70.0 ± 8.1	70.0 ± 8.1	70.0 ± 8.1
DeepHit	500	62.9 ± 11.6	63.0 ± 11.7	63.0 ± 11.6	63.0 ± 11.6	63.0 ± 11.6
Logistic Hazard	500	61.8 ± 11.6	61.8 ± 11.6	61.8 ± 11.6	61.8 ± 11.6	61.8 ± 11.6
N-MTLR	500	69.0 ± 12.6	69.0 ± 12.7	69.0 ± 12.6	69.0 ± 12.6	69.0 ± 12.6
DeepHit	1000	63.3 ± 9.4	63.3 ± 9.4	63.3 ± 9.4	63.3 ± 9.4	63.3 ± 9.4
Logistic Hazard	1000	57.9 ± 13.0	57.9 ± 13.0	57.9 ± 13.0	57.9 ± 13.0	57.9 ± 13.0
N-MTLR	1000	67.3 ± 13.6	67.3 ± 13.6	67.3 ± 13.6	67.3 ± 13.6	67.3 ± 13.6