Estimating Calibrated Individualized Survival Curves with Deep Learning

Fahad Kamran, Jenna Wiens
Computer Science and Engineering
University of Michigan, Ann Arbor, MI
fhdkmrn, wiensj@umich.edu

Abstract

In survival analysis, deep learning approaches have been proposed for estimating an individual's probability of survival over some time horizon. Such approaches can capture complex non-linear relationships, without relying on restrictive assumptions regarding the relationship between an individual's characteristics and their underlying survival process. To date, however, these methods have focused primarily on optimizing discriminative performance and have ignored model calibration. Well-calibrated survival curves present realistic and meaningful probabilistic estimates of the true underlying survival process for an individual. However, due to the lack of ground-truth regarding the underlying stochastic process of survival for an individual, optimizing and measuring calibration in survival analysis is an inherently difficult task. In this work, we i) highlight the shortcomings of existing approaches in terms of calibration and ii) propose a new training scheme for optimizing deep survival analysis models that maximizes discriminative performance, subject to good calibration. Compared to state-of-the-art approaches across two publicly available datasets, our proposed training scheme leads to significant improvements in calibration, while maintaining good discriminative performance.

Introduction

In survival analysis, one aims to learn the relationship between an individual's covariates and the underlying stochastic process of some event (*e.g.*, disease onset). Beyond discriminative performance (*i.e.*, how the relative predictions between individuals match the observed outcomes), to be useful for real-world applications, survival models must be well calibrated (Gneiting and Katzfuss 2014). In clinical settings, making decisions at a patient-level requires survival estimates that are accurate with respect to the ground-truth survival probability. Poor calibration can lead to misleading predictions, resulting in potentially clinically harmful models (Van Calster and Vickers 2015; Van Calster et al. 2019; Shah, Steyerberg, and Kent 2018; Steyerberg et al. 2019). Accurate estimates of survival at different time-points can help augment clinical decision making at a per-patient level.

We define a calibrated model as one that consistently produces estimates of survival that match the underlying survival probabilities for each individual (Haider et al. 2020). To better understand what these individualized underlying survival probabilities represent, consider building an estimate of survival for an entire population using a simple counting-based Kaplan-Meier estimate (Kaplan and Meier 1958). Differences among individuals will lead to events at different time-points, resulting in a decreasing estimate of population-level survival over time. This estimate reflects the variation in the time-to-event distribution. Now consider a set of individuals with identical or near-identical covariates. Despite the similarity among individuals, we might still expect the time-to-event distribution for this homogeneous population to exhibit some variation due to the stochasticity in nature, resulting in a gradually decreasing survival curve for these individuals. Along these lines, the underlying survival curve for an individual should reflect such variation.

Figure 1 illustrates potential differences in discriminative performance and calibration performance via a hypothetical example. The *solid* curves represent the true underlying survival distributions, and the *dashed* lines represent hypothetical estimates for three different individuals. With respect to the observed event times, all three sets of estimated survival curves *correctly rank* the individuals, and hence, have good *discriminative* performance. However, the first two sets of survival curves (a and b) consistently underestimate or overestimate the survivival probabilities with respect to the true survival curve. Hence, these estimates are *miscalibrated*. Meanwhile, the third set of estimated survival curves (c) is well calibrated, since it aligns with the true survival probabilities. These calibrated estimates provide an accurate probabilistic interpretation of survival for an individual throughout the time horizon.

Deep survival models have achieved state-of-the-art discriminative performance by relaxing any distributional assumptions and directly estimating the underlying process (Lee et al. 2018; Ren et al. 2019). However, to date, such models are trained by optimizing for discriminative performance and have not been evaluated in terms of calibration. Though useful for ranking individuals, the resulting survival curves may consistently overestimate or underestimate an individual's probability of survival, as in **Figure 1**.

In light of these issues, we focus on approaches for training and evaluating deep survival analysis models that account for *both* calibration and discriminative performance. Our contributions include:

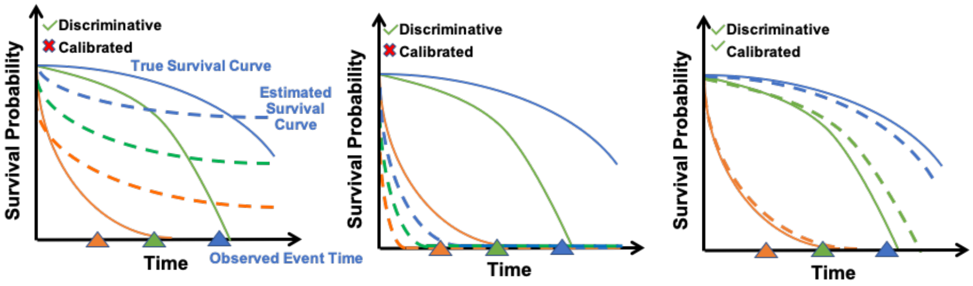

Figure 1: Hypothetical Example. Three hypothetical sets of estimated survival curves for three individuals (dashed) and their corresponding true underlying survival distributions (solid), where the triangles represent the observed event times. All three sets of estimated curves correctly rank the individuals (*i.e.*, have good discriminative performance). However, the first two sets of estimated survival curves consistently overestimate or underestimate the true survival probability at various points throughout the time horizon. Meanwhile, the third set of estimated survival curves closely aligns with the true survival curves. Hence, the estimated survival curves more accurately reflect the probability of survival. The first two sets of estimated survival curves are *miscalibrated*, while this third set of estimated survival curves are *well-calibrated*.

- we highlight the shortcomings of existing methods for training and evaluating in terms of calibration,
- we propose a novel training scheme for deep survival analysis models and provide theoretical justification for why this training scheme should result in well-calibrated survival estimates, and
- we empirically demonstrate that the proposed training scheme leads to well-calibrated models, while remaining competitive in terms of discriminative performance

We present a framework for training and evaluating deep survival models that focuses on calibration. Through a series of experiments on two publicly available datasets, we compare our approach to state-of-the-art approaches in survival analysis, demonstrating the proposed approach's ability to maximize discrimination subject to good calibration.

Background and Related Work

Here, we formalize the core survival analysis problem and introduce notation. We then survey training schemes in deep survival analysis that have achieved state-of-the-art performance, while not relying on specific assumptions regarding the distributional form of the relationship between an individual's covariates and their survival probability.

Problem Setup and Notation

Survival analysis aims to learn a time-to-event model using data of the form $D = \{(\mathbf{x}_i, z_i, c_i)\}_{i=1}^{n}$, where n is the total number of individuals. Each $(\mathbf{x}_i, z_i, c_i) \in D$ represents information for one individual, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the individual's covariates, z_i denotes the observed time of the event, or time of censoring, and c_i denotes the individual's censoring status. In this work, we only consider right-censoring, the most common scenario in survival analysis (Cox 1972; Kaplan and Meier 1958; Shivaswamy, Chu, and Jansche 2007; Wang, Li, and Reddy 2019). An individual i is said to be right-censored ($c_i = 1$) if the event did not occur at time z_i, but instead, the individual was lost to follow-up (*i.e.*, censored) after this time.

In this work, we aim to accurately estimate individualized survival probabilities over some discrete time horizon (Haider et al. 2020; Lee et al. 2018). Given data from D, our goal is to learn a model f that maps covariates for individual i $\mathbf{x}_i$ to *individualized* estimates of $P(Z = t|\mathbf{x}_i)$ for $t \in \{0, 1, ..., \tau\}$, where time is binned into τ intervals (Lee et al. 2018; Ren et al. 2019). From these estimates, we can estimate the survival curves $S(t|\mathbf{x}_i) = P(Z > t|\mathbf{x}_i) = \sum_{j>t} P(Z = j|\mathbf{x}_i)$ and the cumulative incidence function (CIF) $F(t|\mathbf{x}_i) = P(t \leq Z|\mathbf{x}_i) = \sum_{j \leq t} P(Z = j|x_i)$

Achieving good discriminative performance means accurately rank at-risk individuals. Formally, for any two individuals with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$, assume individual 1 has the event at time z_1, at which individual 2 has not had the event nor have they been censored (*i.e.*, $z_2 > z_1, c_1 = 0, c_2 \in \{0, 1\}$). Then, we would expect individual 1 to be at greater risk than individual 2 at time z_1, or $\hat{F}(z_1|\mathbf{x}_1) > \hat{F}(z_1|\mathbf{x}_2)$. This is often measured through the C-index, which calculates the proportion of unique pairs of individuals (that match the criteria above) for which this ranking is correct (Antolini, Boracchi, and Biganzoli 2005; Lee et al. 2018).

Well-calibrated models should produce survival estimates $\hat{S}(\cdot|\mathbf{x}_i)$ that match the underlying survival distribution $S(\cdot|\mathbf{x}_i)$. The Brier score, defined at time t as $\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{t \leq z_i} - \hat{S}(t|\mathbf{x}_i))^2$, is often used to measure calibration (Murphy 1973; Lee et al. 2019; Kvamme, Borgan, and Scheel 2019). However, the Brier score measures how well a prediction matches the observed outcome for different individuals, which differs from the definition of calibration considered here. In particular, a discontinuous heaviside step function that equals 0 at and after the observed event time could qualify as perfectly calibrated as it perfectly matches the observed outcome (*i.e.*, average Brier score = 0), despite no meaningful probabilistic interpretation (i.e., it does not correctly reflect the variation in the probability estimate due to stochasticity in nature). Moreover, the Brier score over the full survival curve is heavily influenced by the choice of

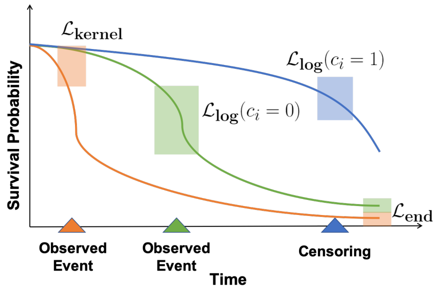
$\mathcal{L}_{\text{kernel}}$
$\mathcal{L}_{\text{log}}(c_i = 1)$
$\mathcal{L}_{\text{log}}(c_i = 0)$
$\mathcal{L}_{\text{end}}$
Survival Probability
Observed Event
Observed Event
Censoring
Time

Figure 2: Each loss function provides a different kind of supervision throughout the time horizon (shaded region), but none explicitly focuses on calibration.

time-horizon. Accordingly, the Brier score is insufficient as an evaluation metric in our setting. In this work, we explore other ways to measure calibration that fit our definition, such as D-Calibration (Andres et al. 2018; Haider et al. 2020).

Deep Survival Analysis Training Schemes

The choice of the training scheme used to optimize a deep survival analysis model defines its success in terms of both discriminative performance and calibration. Common objective functions include:

- $\mathcal{L}_{log} = -\sum_{i=1}^{n}(1 - c_i) \cdot \log(\hat{P}(Z = z_i|\mathbf{x}_i)) + c_i \cdot \log(\hat{S}(z_i|\mathbf{x}_i))$
- $\mathcal{L}_{end} = -\sum_{i=1}^{n}(1 - c_i) \cdot \log(1 - \hat{S}(\tau|\mathbf{x}_i))$
- $\mathcal{L}_{kernel} = \sum_{i \neq j} A_{i,j} \cdot exp(\frac{-(\hat{S}(z_i|\mathbf{x}_j) - \hat{S}(z_i|\mathbf{x}_i))}{\sigma})$, where $A_{i,j} = \mathbb{1}_{c_i = c_j = 0, z_i < z_j}$

$\mathcal{L}_{log}$, often termed the logarithmic loss, maximizes the estimated probability of the event occurring at the time of observation, while maximizing the estimated survival probability at the time of censoring for censored individuals (Lee et al. 2018; Ren et al. 2019). $\mathcal{L}_{end}$, often used in conjunction with $\mathcal{L}_{log}$, adds supervision after the observed event time, by forcing the survival probability to zero at the final timestep for uncensored individuals (Ren et al. 2019). Lastly, $\mathcal{L}_{kernel}$ penalizes incorrectly ordering two uncensored individuals (Lee et al. 2018). **Figure 2** shows where these different loss functions provide supervision over the time horizon. Most deep survival models use $\mathcal{L}_{log}$ during training (Miscouridou et al. 2018; Lee et al. 2018; Ren et al. 2019). $\mathcal{L}_{kernel}$ was explored in early deep survival analysis works as a method for increasing discriminative performance, but has been less explored recently (Lee et al. 2018). Recent state-of-the-art has shown strong discriminative performance when training using a composite of $\mathcal{L}_{log}$ and $\mathcal{L}_{end}$ (Ren et al. 2019).

Though the logarithmic loss corresponds to a proper scoring rule, it is sensitive to extreme cases and outliers (Gneiting and Raftery 2007; Gneiting and Katzfuss 2014). This sensitivity results in a larger trade-off between making accurate predictions and maintaining calibration compared to other proper scoring rules, such as the continuous-rank probability score (CRPS). These methods have not been evaluated for their calibration performance. We hypothesize that the models trained to minimize $\mathcal{L}_{log}$ could result in miscalibrated survival estimates. In light of this observation, we consider loss functions that build off of proper scoring rules without this limitation. In particular, our proposed approach builds on the CRPS, which is defined as $\int_{-\infty}^{\infty}(\hat{F}(t|\mathbf{x}_i) - \mathbb{1}_{z \leq t})^2 dt$, which has been explored in survival analysis (Avati et al. 2020). However, this objective function relies on an infinite integral and thus requires specific distributional assumptions during training. In contrast, our discrete approximation avoids relying on any distributional assumptions. Moreover, we consider how this discrete approximation can be incorporated into a training scheme with other loss functions to elicit calibrated and accurate survival estimates. Finally, we consider a comprehensive evaluation framework for properly measuring the efficacy of survival models for both their discriminative performance and calibration. Concurrent work to ours proposed directly optimizing for a variant of a calibration metric we use for evaluation (Goldstein et al. 2020). Future work might consider how the two proposed training schemes could be combined for further improvements.

Methods

In this section, we present our proposed training scheme and our comprehensive evaluation metrics. We begin by proposing a new loss function and theoretically justifying why it should elicit survival models with good discriminative performance and good calibration. We continue by discussing and justifying our proposed training scheme, which consists of combining this new loss function with $\mathcal{L}_{kernel}$. We explain why this combination should improve both overall performance. We conclude with a discussion on how to evaluate models for both discriminative performance and calibration.

Proposed Training Scheme

We propose minimizing the rank probability score (RPS), $\mathcal{L}_{RPS}$, defined as:

$$\sum_{i=1}^{n}(1 - c_i) \cdot \sum_{t=1}^{\tau}(\hat{S}(t|\mathbf{x}_i) - \mathbb{1}_{t < z_i})^2 + c_i \cdot \sum_{t=1}^{z_i}(\hat{S}(t|\mathbf{x}_i) - 1)^2$$

$\mathcal{L}_{RPS}$ focuses on the relevant portions of the full time-horizon, rather than just the specific event-time. For uncensored individuals ($c_i = 0$), $\mathcal{L}_{RPS}$ pushes the survival probability at times before an individual has an event to 1, and shrinks the survival probability to 0 at times after the event has occurred. For uncensored individuals, $\mathcal{L}_{RPS}$ is averaged over the full time horizon τ, as we have access to the survival status for the full time interval. For censored individuals ($c_i = 1$), $\mathcal{L}_{RPS}$ pushes the survival probability to 1 before the individual is censored, and is averaged over the available time horizon for censored individuals z_i, as we do not know their survival status after this time.

Claim. *Training deep survival models using $\mathcal{L}_{RPS}$ will result in well-calibrated estimates of survival.*

Proof. Consider n individuals with identical or near-identical covariates with observed event times $\{z_i\}_{i=1}^{n}$. Define the counting-based Kaplain-Meier estimate for these individuals at time t as $KM_t^n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{t < z_i}$, where

$\lim_{n\to\infty} KM_t^n$ is the underlying survival probability at time t for these n individuals.

A survival model will estimate one survival probability for these n individuals at time t. Define this value as $\hat{p}_t$. A well-calibrated survival model will output a $\hat{p}_t$ that closely aligns with the underlying survival probability $\lim_{n\to\infty} KM_t^n$. Consider the optimization problem of finding $\hat{p}_t$ which will minimize $\mathcal{L}_{RPS}$. This problem can formally be set-up as $\arg\min_{\hat{p}_t} \sum_{i=1}^{n}(\hat{p}_t - \mathbb{1}_{t<z_i})^2$.

First, this optimization problem is strictly convex and has a unique minimum, as the second derivative is positive everywhere (see Supplementary Material).

To find the value of $\hat{p}_t$ that minimizes this objective function ($\hat{p}_t^*$), we set the derivative equal to zero.

$$\frac{\partial}{\partial \hat{p}_t^*}\left(\sum_{i=1}^{n}(\hat{p}_t^* - \mathbb{1}_{t<z_i})^2\right) = 0$$

$$2\hat{p}_t^* - \frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{t<z_i} = 0$$

$$\hat{p}_t^* = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{t<z_i}$$

The unique estimated survival probability that minimizes the objective function is equivalent to the average survival status for all n individuals at time t. This unique minimum is equal to KM_t^n which, as n gets large, is equal to the true underlying survival probability for these individuals at time t. Hence, training a survival model to minimize $\mathcal{L}_{RPS}$ will result in estimated survival probabilities that align well with the true survival probabilities. □

A model that minimizes $\mathcal{L}_{RPS}$ will theoretically result in well-calibrated survival estimates that align well with the true survival curves. However, due to the inherent noise in the training process of deep models and the inability to guarantee a global solution, training using just $\mathcal{L}_{RPS}$ as a loss function might be insufficient. In particular, combining $\mathcal{L}_{RPS}$ with a loss function that can scale survival probabilities and encourages good discriminative ability would improve overall performance.

Hypothesis. *Training deep survival models using a composite loss function $\mathcal{L}_{RPS} + \lambda\mathcal{L}_{kernel}$, yields an accurate, yet calibrated survival model when the value of σ in $\mathcal{L}_{kernel}$ is appropriately tuned.*

Justification. Remember that $\mathcal{L}_{kernel}$ is defined as $\mathcal{L}_{kernel} = \sum_{i\neq j} A_{i,j} \cdot exp(\frac{-(\hat{S}(z_i|\mathbf{x}_j)-\hat{S}(z_i|\mathbf{x}_i))}{\sigma})$. In this loss function, σ controls the scale of the differences between survival probabilities for different individuals. When σ is small (*i.e.* $\sigma \leq .1$) and individuals are correctly ranked, small or large differences between two individual's survival probabilities (numerator) minimize $\mathcal{L}_{kernel}$. In contrast, when σ is large (*i.e.* $\sigma \geq 10$) and individuals are correctly ranked, only large differences between individual's survival probabilities can minimize $\mathcal{L}_{kernel}$. Hence, the value of σ can directly affect how the variation of different individuals survival curves over the interval $[0, 1]$. In particular, we expect that training a model to minimize $\mathcal{L}_{kernel}$ with a small σ value will result in survival curves that are not well-spread out, while training a model to minimize $\mathcal{L}_{kernel}$ with a large σ value will scale the survival curves in order to spread them out sufficiently. The value of σ should be tuned based on a validation set.

The ability to control the variation of individual's survival curves can also be thought of as rescaling survival curves in order to best minimize $\mathcal{L}_{kernel}$. If $\mathcal{L}_{RPS}$ overestimates or underestimates the survival probability for individuals at certain times, using $\mathcal{L}_{kernel}$ with an appropriately tuned value of σ can scale these estimates to more accurately estimate the true underlying survival probabilities. At the same time, as $\mathcal{L}_{kernel}$ aims to correctly rank individuals, it will still maximize discriminative performance. Thus, we expect that the combination of $\mathcal{L}_{kernel}$ and $\mathcal{L}_{RPS}$ will encourage good calibration without sacrificing discrimination.

The value of λ helps control the trade-off between the two loss functions in the composite loss. As setting λ to 0 translates to simply the $\mathcal{L}_{RPS}$ loss function, and setting λ too high translates to the $\mathcal{L}_{kernel}$ loss function, we hypothesize that an intermediate value of λ will result the best trade-off between the theoretical guarantees of correctly estimating the underlying survival probability obtained by minimizing $\mathcal{L}_{RPS}$ and the scaling ability of $\mathcal{L}_{kernel}$.

In summary, we introduced a novel loss function $\mathcal{L}_{RPS}$, which we hypothesize will result in increased calibration performance when used to train survival analysis models. Moreover, we proposed a new training scheme that involves minimizing a composite loss of $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$.

Evaluating Model Performance

We evaluate model performance in terms of both discrimination and calibration. We evaluate discriminative performance, in terms of the aforementioned C-index, which calculates the proportion of individuals who are correctly ranked by the estimated models. To measure calibration, we consider the average Brier score (*i.e.*, mean-square-error over the survival curve) and D-Calibration (Haider et al. 2020; Andres et al. 2018). Brier score measures how well a prediction matches the observed outcome for different individuals, and hence, does not fully capture our definition of calibration. D-Calibration bins the estimated survival probabilities at the true event times into ten equal-width intervals between 0 and 1, and performs a chi-squared test to determine if the distribution is uniform. This more closely aligns with our definition of calibration; however, the test assumes the model is well-calibrated, placing the burden on disproving the null hypothesis.

In light of these shortcomings, we also consider the **distributional divergence for calibration (DDC)**. DDC does not rely on a statistical test and produces a continuous score that allows for comparisons of different models. Given a set of estimated survival probabilities for each individual at their observed event times $\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^{n}$, we compute DDC as the Kullback-Leibler (KL) Divergence $D_{KL}(P||Q)$ between a binned distribution $P = B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^{n})$ and the uniform distribution Q, where B is a function that maps a set of probabilities into a probability distribution over $\mathcal{X}$, ten equal-width bins covering the unit interval (Lin 1991). Due to the discrete nature of the binning operation, we change the base of the logarithm when calculating DDC to ensure that it ranges between 0 and 1. DDC measures the distance between the empirical distribution of estimated probability of survival at the time of the events P and the uniform distribution Q. Lower is better; if $P = Q$, then $DDC(P, Q) = 0$. Survival curves that estimate a single survival probability, such as 0, for every individual at their observed event time, for which $B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n) = B(\{0\}^n)$, achieve a maximum DDC of 1.

Claim. *A perfectly calibrated survival model necessarily minimizes the divergence between P and Q for a sufficiently large n.*

Proof. The probability integral transform states that for some random variable X with cumulative distribution function F_x, $F_x(X)$ should be uniformly distributed $U(0, 1)$ (Angus 1994). Thus, given a randomly sampled event time z_i, it must be that $S(z_i) = 1 - F(z_i) \sim U(0, 1)$. Given a set of randomly sampled event times $\{z_i\}_{i=1}^n$, where n is sufficiently large (e.g., $n >> 10$), we then expect the distribution of $P = B(\{\hat{S}(z_i|\mathbf{x}_i)\}_{i=1}^n) \sim U(0, 1)$ (Haider et al. 2020). Hence, a calibrated survival model should minimize the divergence between P and a uniform distribution Q. □

Though necessary, minimizing this metric does not *guarantee* that the estimated survival curves accurately estimate the true underlying survival process. Despite good calibration, these probabilistic estimates may still be inaccurate (i.e., poor discrimination). Hence, it is important to evaluate models in terms of both their calibration and their discriminative performance. To this end, we seek models that excel with respect to *both* measures of performance.

Importantly, DDC does is not applied to censored individuals. Though learning with censored individuals is a key element of survival analysis, evaluating calibration on censored individuals raises a number of issues. Without strong assumptions on the event time distribution for censored individuals, one cannot make meaningful conclusions regarding the calibration of a model for censored individuals (see Supplementary Material for discussion). To this end, while we measure discriminative performance across both uncensored and censored individuals, we focus our evaluation of calibration (specifically, DDC and D-Calibration) on uncensored individuals. This introduces a mismatch between the distribution we evaluate in practice and the one we aim to evaluate in theory. However, if patients are censored at-random, this estimate of calibration should generalize.

Tradeoff between calibration and discrimination. It is important to note that well-calibrated survival curves need not have optimal discriminative performance on the observed sample. A well-calibrated model is one that consistently estimates survival curves that closely match the true survival curves. Due to stochasticity, some individuals may experience the event when their true survival probability is high. As a perfectly calibrated model will estimate a high survival probability at the observed event time for these individuals, these individuals will contribute negatively to the C-Index calculated based on the observed event times. Hence, when individual time to event varies (which we expect is often the case due to the stochasticity of nature), there exists a trade-off between obtaining perfect calibration and perfect observed discriminative performance (i.e. a C-index of 1). This issue arises due to discrimination being measured with respect to only single observed sample. This phenomenon is explored further in the Supplementary Material.

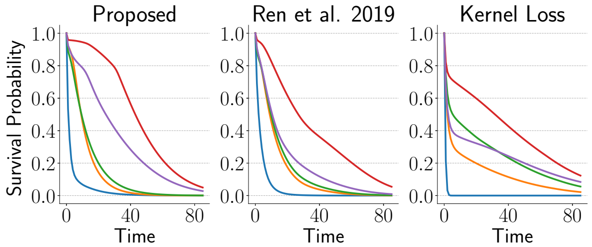

Figure 3: Example survival curves estimated using DRSA trained with $\mathcal{L}_{log} + \mathcal{L}_{end}$ (left), example survival curves estimated using DRSA trained with the proposed training scheme (middle),and example survival curves estimated using DRSA trained with $\mathcal{L}_{kernel}$ (right) on the NACD dataset. Each color represents a randomly selected individual from the test set; the same individuals are shown in each graph. Visually, training with the proposed scheme results in survival curves with a greater variation in shape over time, due to the supervision over the full time horizon and the relative scaling abilities of $\mathcal{L}_{kernel}$.

Practically speaking, both measures of performance are important. Maintaining discriminative performance with increased calibration represents an important gain for a particular survival model. Accordingly, we consider the trade-off between the discriminative performance and calibration by calculating the harmonic mean between the C-index and $1 - DDC$, a value we term the **total score**. A higher total score corresponds to a model that balances discriminative performance and calibration.

Experiments and Results

Here, we test the efficacy of the proposed training scheme. We present two publicly available datasets on which we test our proposed methods and benchmark methods to which we compare. We detail the proposed method's performance compared to the benchmarks in terms of discrimination and calibration and compare against different ablations of the proposed method using the new evaluation framework.

Experimental Setup

Datasets. We consider two publicly available datasets:

- the **Northern Alberta Cancer Dataset (NACD)** consists of 2,402 individuals with various forms of cancer (Haider et al. 2020; Yu et al. 2011). The dataset tracks 51 features for each individual, including demographics, vital signs, patient characteristics such as appetite, and specific details about the type and progression of the cancer. 36.6% of the individuals in the dataset are right-censored, with an average survival time of 16.06 months for uncensored individuals. For this dataset, we use a τ of 86 months based on the largest length of stay.

- **CLINIC** records the survival status of 6,036 patients in a hospital, with 13.2% being censored (Knaus et al. 1995). The dataset consists of 14 features for each individual, including information about demographics, vital signs, onset of diseases, and medications. The average survival time for uncensored individuals is 5.33 months. For this dataset, we use a τ of 52 months, such that each time-bin represents one month.

Model Architecture. To demonstrate the efficacy of our approach and compare it against baseline methods, we consider minimizing the proposed composite loss to train the Deep Recurrent Survival Analysis (DRSA) architecture (Ren et al. 2019). Though the proposed loss functions are model-agnostic, we consider the DRSA architecture due to its state-of-the-art discriminative performance, flexibility for allowing variable-length forecasting, and its lack of assumptions regarding the probability at the end of the time-horizon. More information about this architecture choice can be found in the Supplementary Material.

Baselines. To evaluate how our proposed approach compares to current state-of-the-art in deep survival analysis, we compare against two baseline survival analysis models:

- The DRSA architecture with the objectives it was originally proposed with (using $\mathcal{L}_{log}$ and $\mathcal{L}_{end}$) (Ren et al. 2019), and
- Multi-task logistic regression (MTLR) is one of the only survival analysis approaches that has shown good empirical performance in terms of our definition of calibration (Yu et al. 2011; Haider et al. 2020). MTLR trains a separate logistic regression model per time-point to estimate survival, and combines these to estimate the survival distribution over some time horizon. When compared to other methods, such as extensions of the Cox model, MTLR performed best in terms of both calibration and discrimination (Haider et al. 2020).

Training/Evaluation Details. Across experiments, we use the same DRSA architecture: a one-layer LSTM with hidden size 100 and a single feed-forward layer with a sigmoid activation on the output for each time-step (Ren et al. 2019). We separate our data into training/validation/test sets using a 60/20/20% split. For training, we use Adam and a batch size of 50 (Kingma and Ba 2015). We train for 100 epochs (which, empirically, was enough for models to converge) and select the best model based on a validation set. For the proposed composite training scheme, we tune the value of σ for $\mathcal{L}_{kernel}$ based on the NACD dataset, and use this optimal value on the CLINIC dataset to test whether the manner in which $\mathcal{L}_{kernel}$ affects $\mathcal{L}_{RPS}$ generalizes across multiple datasets. When training with multiple losses, we use $\lambda = 1$. Though we considered other weighting schemes, it did not appear to affect performance. Note that we weighted the $\mathcal{L}_{RPS}$ loss function due to the right-skewed time-to-event distribution for both datasets. We train each model five times, with different weight initializations. We present the mean and the standard deviation of the results on the test set for all metrics except D-Calibration, for which we present the number of runs where the resulting survival estimates passed the D-Calibration test. We evaluate DDC and D-calibration using only uncensored test individuals, but we evaluate C-index and Brier score using all test individuals. All deep models were built in PyTorch [1], while MTLR was implemented using the corresponding R package (Paszke et al. 2019; Haider 2019).

Results

First, our proposed approach consistently outperforms all baselines with respect to DDC and D-calibration, while maintaining comparable C-index and average Brier score values (**Table 1**). Lower values represent better performance for DDC and Brier score, while higher values represent better performance for the other metrics. The proposed method consistently leads to estimated survival curves with a better trade-off between calibration and discrimination, as evidenced by the higher total score compared to MTLR and DRSA as it was originally proposed. The fact that no model dominates in C-index across datasets is consistent with recent findings in survival analysis (Lee et al. 2019).

Compared to the original DRSA (Ren et al. 2019), the proposed training scheme results in a statistically significant improvement in calibration across both tasks (NACD DDC: .025 vs. .007, CLINIC DDC: .138 vs .056). This improvement, however, is accompanied by a small decrease in C-index in the NACD dataset. However, the probabilistic estimates of survival are more likely to accurately represent the true underlying survival processes. We see the same overall trend when comparing our proposed method with MTLR, where the proposed model is significantly more calibrated across both datasets (NACD DDC: .062 vs .007, CLINIC DDC: .168 vs .057), while the relative C-index depends on the dataset.

Compared to training each component of the proposed loss (*i.e.*, $\mathcal{L}_{RPS}$ and $\mathcal{L}_{kernel}$) separately, using the composite loss leads to improvements (**Table 1**: NACD total score: .715 and .847 vs .850, CLINIC total score: .687 and .731 vs .753). In particular, note that training with $\mathcal{L}_{RPS}$ results in good calibration performance, while training with $\mathcal{L}_{kernel}$ in and of itself results in poor calibration performance. Hence, as expected, $\mathcal{L}_{RPS}$ itself will elicit calibrated and accurate estimates of survival, but combining it with the scaling ability of $\mathcal{L}_{kernel}$ can improve performance even more. Moreover, training using $\mathcal{L}_{RPS}$ alone results in better calibration than using the logarithmic loss functions (NACD DDC: .025 vs .012, CLINIC DDC: .138 vs .097), with minimal drops in discriminative performance. These empirical results support the original hypothesis that training using $\mathcal{L}_{RPS}$ should result in survival models that better balance discriminative performance and calibration, but the composite loss results in the best performance.

Next, we focus on a qualitative assessment of our proposed method. Visually, this approach produces survival curves with a greater variation in shape over the full time horizon (**Figure 3**). In particular, the baseline training scheme results in survival curves that decay quickly towards a survival probability of 0. This is evidenced by the high DDC value due to many individuals' estimated survival

[1] https://github.com/MLD3/Calibrated-Survival-Analysis

Model	**NACD**					**CLINIC**				
	C-index ↑	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**	**C-index ↑**	**DDC ↓**	**D-Calibration ↑**	**$\overline{\text{Brier}}$ ↓**	**Total Score ↑**
Ren et al. 2019	.748 ± .002	.025 ± .012	1	.101 ± .002	.846 ± .004	.616 ± .003	.138 ± .002	0	.107 ± .000	.719 ± .003
MTLR	.750 ± .000	.062 ± .000	0	.101 ± .000	.834 ± .000	.608 ± .000	.168 ± .000	0	.106 ± .000	.702 ± .000
Proposed - $\mathcal{L}_{RPS}$	.741 ± .008	.305 ± .089	0	.207 ± .034	.715 ± .050	.628 ± .003	.241 ± .022	0	.153 ± .002	.687 ± .011
Proposed - $\mathcal{L}_{kernel}$	.742 ± .003	.012 ± .002	3	.101 ± .003	.847 ± .001	.615 ± .005	.097 ± .006	0	.110 ± .001	.731 ± .005
Proposed	.742 ± .006	**.007 ± .003***	5	.104 ± .002	.850 ± .003	.627 ± .001	**.056 ± .011***	0	.106 ± .001	**.753 ± .004***

Table 1: The proposed training approach consistently leads to improvements in calibration (DDC, D-Calibration, Averaged Brier Score) across all baselines and ablations, without sacrificing discriminative performance (C-index) (mean ± standard deviation across random initializations, number of times passing the statistical test for D-Calibration). Lower DDC and Brier scores and higher values of C-index, D-Calibration, and total score indicate better performance. An * indicates results that are statistically significant over all baselines using a paired t-test ($p < .05$).

probabilities being very low at the time they experienced the event. Meanwhile, our proposed loss functions achieve better DDC values by allowing more flexibility in the shape of the survival curves, such that some individuals have higher survival probabilities at the time of their observed events. We hypothesize that this is due in part to the direct supervision over the entire predictive distribution that comes from training with $\mathcal{L}_{RPS}$. In contrast, $\mathcal{L}_{log}$ provides direct supervision on the survival probability over only a single time-point, possibly resulting in less flexibility in the shape of the predictive distribution over the time horizon (Gneiting, Balabdaoui, and Raftery 2007). This single time-point supervision, along with the logarithmic losses sensitivity to extreme cases, can result in miscalibrated survival curves.

We present results for proposed method using $\sigma = 0.8$ in $\mathcal{L}_{kernel}$ for both datasets. This value was tuned on a validation set on the NACD dataset and applied to the CLINIC dataset. Hence, the manner in which $\mathcal{L}_{kernel}$ affects $\mathcal{L}_{RPS}$ generalizes across multiple datasets, supporting our original hypothesis. Moreover, we visually confirm the original motivation for the use of $\mathcal{L}_{kernel}$: the value of σ helps control the scale of different individual's survival curves. As noted in Section 3, we expect a model trained to minimize $\mathcal{L}_{kernel}$ with small σ (*e.g.* σ = 0.1) to result in survival curves where different individuals curves are close to each other in scale, and a model trained to minimize $\mathcal{L}_{kernel}$ with large σ (*e.g.* σ = 10.0) to result in more spaced out survival curves. **Figure 4** shows estimated example curves for 10 random individuals in the NACD dataset when trained using $\mathcal{L}_{kernel}$ with σs of 0.1 and 10.0. The resulting survival curves display the hypothesized phenomenon, confirming the ability of $\mathcal{L}_{kernel}$ to control the scale of different individuals' survival curves. Hence, the improved performance for the composite loss is in part due to an additional rescaling of the survival distributions to better match the underlying survival probabilities.

Overall, these results indicate the ability of our proposed training procedure to better match the true survival distribution, while maintaining the useful property of accurately ranking individuals. Moreover, the comprehensive evaluation framework helps facilitate model comparisons for both discriminative performance and calibration.

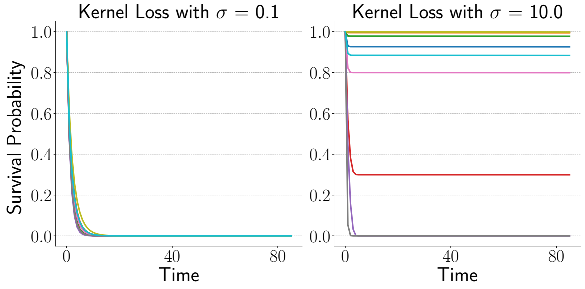

Figure 4: Survival curves from models trained with $\mathcal{L}_{kernel}$ using $\sigma = 0.1$ (left) and $\sigma = 10.0$ (right) from the NACD dataset. Each color represents a different individual. These plots confirm our original hypothesis regarding $\mathcal{L}_{kernel}$: the value of σ can control the relative scales of survival probabilities. Hence, by tuning σ, we can scale the survival curves to best match the true underlying survival distributions.

Conclusion

Given the stochasticity of nature, we expect individuals to have an *underlying survival distribution* that corresponds to a meaningful probabilistic interpretation of an individual's survival. Though critical to clinical application, calibration to date has been largely overlooked in survival analysis, especially in deep survival analysis. We hypothesized that recent work in deep survival analysis that optimizes and evaluates for discriminative performance alone results in poorly-calibrated estimated survival curves. To this end, we introduced a new approach for training deep survival analysis models to optimize for both discriminative performance and calibration. We provided both theoretical justification and empirical evidence for why the proposed approach elicits calibrated estimates of survival. Applied in the context of a state-of-the-art deep survival analysis architecture, the proposed training scheme leads to significant gains in calibration across two publicly available datasets, while achieving similar discriminative performance. Still, there remains room for improvement. In particular handling continuous-time survival analysis problems without the use of any distributional assumptions is an interesting line of future work. Nonetheless, this work presents a complete and flexible pipeline for training and evaluating accurate and well-calibrated deep models for survival analysis.

Acknowledgments

This project was supported by the National Science Foundation Career award (IIS-1553146). We would also like to thank the anonymous reviewers for their valuable feedback. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

Ethics Statement

We propose a comprehensive training and evaluation scheme for building accurate and calibrated individualized survival distributions. The contributions in this work can have broad societal impacts, especially in the context of healthcare. In particular, our work provides major contributions towards the goal of personalized healthcare by focusing on estimating individualized survival curves. Our definition of calibration ensures an emphasis on estimating the true underlying survival distribution for a particular individual. Accurate and calibrated individualized survival distributions can help augment clinical decision making on a per-patient basis, rather than at a population level, providing important steps towards personalized medicine. However, this is only the first step. Going forward, survival analysis techniques, such as those proposed in this work, should be combined with recent advances in fairness in machine learning to ensure accurate and useful personalized algorithms that do not reinforce harmful biases present in the original data.

References

Andres, A.; Montano-Loza, A.; Greiner, R.; Uhlich, M.; Jin, P.; Hoehn, B.; Bigam, D.; Shapiro, J. A. M.; and Kneteman, N. M. 2018. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS one* 13(3): e0193523.

Angus, J. E. 1994. The probability integral transform and related results. *SIAM review* 36(4): 652–654.

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

Avati, A.; Duan, T.; Zhou, S.; Jung, K.; Shah, N. H.; and Ng, A. Y. 2020. Countdown regression: sharp and calibrated survival predictions. In *Uncertainty in Artificial Intelligence*, 145–155. PMLR.

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

Gneiting, T.; Balabdaoui, F.; and Raftery, A. E. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2): 243–268.

Gneiting, T.; and Katzfuss, M. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* 1: 125–151.

Gneiting, T.; and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477): 359–378.

Goldstein, M.; Han, X.; Puli, A.; Perotte, A.; and Ranganath, R. 2020. X-CAL: Explicit Calibration for Survival Analysis. *Advances in Neural Information Processing Systems* 33.

Haider, H. 2019. *MTLR: Survival Prediction with Multi-Task Logistic Regression*. URL https://CRAN.R-project.org/package=MTLR. R package version 0.2.1.

Haider, H.; Hoehn, B.; Davis, S.; and Greiner, R. 2020. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research* 21(85): 1–63.

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

Kingma, D. P.; and Ba, J. 2015. Adam: a method for stochastic optimization. ICLR (2015).

Knaus, W. A.; Harrell, F. E.; Lynn, J.; Goldman, L.; Phillips, R. S.; Connors, A. F.; Dawson, N. V.; Fulkerson, W. J.; Califf, R. M.; Desbiens, N.; et al. 1995. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine* 122(3): 191–203.

Kvamme, H.; Borgan, Ø.; and Scheel, I. 2019. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research* 20(129): 1–30.

Lee, C.; Zame, W.; Alaa, A.; and Schaar, M. 2019. Temporal Quilting for Survival Analysis. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 596–605.

Lee, C.; Zame, W. R.; Yoon, J.; and van der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37(1): 145–151.

Miscouridou, X.; Perotte, A.; Elhadad, N.; and Ranganath, R. 2018. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, 244–256.

Murphy, A. H. 1973. A new vector partition of the probability score. *Journal of applied Meteorology* 12(4): 595–600.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ren, K.; Qin, J.; Zheng, L.; Yang, Z.; Zhang, W.; Qiu, L.; and Yu, Y. 2019. Deep Recurrent Survival Analysis. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Shah, N. D.; Steyerberg, E. W.; and Kent, D. M. 2018. Big data and predictive analytics: recalibrating expectations. *Jama* 320(1): 27–28.

Shivaswamy, P. K.; Chu, W.; and Jansche, M. 2007. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 655–660. IEEE.

Steyerberg, E. W.; et al. 2019. *Clinical prediction models*. Springer.

Van Calster, B.; McLernon, D. J.; Van Smeden, M.; Wynants, L.; and Steyerberg, E. W. 2019. Calibration: the Achilles heel of predictive analytics. *BMC medicine* 17(1): 1–7.

Van Calster, B.; and Vickers, A. J. 2015. Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making* 35(2): 162–169.

Wang, P.; Li, Y.; and Reddy, C. K. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)* 51(6): 110.

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