Countdown Regression: Sharp and Calibrated Survival Predictions

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H. Shah, Andrew Y. Ng
Stanford University
{avati,tonyduan,sharonz,ang}@cs.stanford.edu,{kjung,nigam}@stanford.edu

Abstract

Probabilistic survival predictions from models trained with Maximum Likelihood Estimation (MLE) can have high, and sometimes unacceptably high variance. The field of meteorology, where the paradigm of maximizing sharpness subject to calibration is popular, has addressed this problem by using scoring rules beyond MLE, such as the Continuous Ranked Probability Score (CRPS). In this paper we present the *Survival-CRPS*, a generalization of the CRPS to the survival prediction setting, with right-censored and interval-censored variants. We evaluate our ideas on the mortality prediction task using two different Electronic Health Record (EHR) data sets (STARR and MIMIC-III) covering millions of patients, with suitable deep neural network architectures: a Recurrent Neural Network (RNN) for STARR and a Fully Connected Network (FCN) for MIMIC-III. We compare results between the two scoring rules while keeping the network architecture and data fixed, and show that models trained with Survival-CRPS result in sharper predictive distributions compared to those trained by MLE, while still maintaining calibration.

1 INTRODUCTION

Accurate and confident predictions of the time to an event, such as patient mortality or customer churn, allow for better decision making. Methods have been developed in the survival analysis literature to address this problem of predicting time to events given censored data; that is, when sometimes we only know that an event did not happen until a certain period, or when we know an event happened in a (wide) time window, but not the exact moment. The most common approach for fitting survival models is via Maximum Likelihood Estimation (MLE) [1], or maximum partial likelihood estimation such as in Cox Regression [2]. However, MLE is equivalent to the logarithmic scoring rule, which is known to be subject to hypersensitivity [3, 4]. The intuition is that forecasts are encouraged to be under-confident since the observation of an event that has low predicted density results in an extremely high loss. This can result in models that are overly conservative, and predict distributions with very high variance – sometimes too high to be practically useful. Hypersensitivity of MLE is generally not a problem in typical homoskedastic regression tasks where the models only output point estimates, and uniform variance is assumed across examples. It is also not a problem in applications where the goal is accurate ranking (for example, in risk stratification). However, in the heteroskedastic regression task, where the model outputs a full probability distribution (such as a patient specific survival curve) over the outcome, the shortcomings of MLE can become a problem for practical use.

Having access to accurate and confident probabilistic predictions can especially be helpful in healthcare. Historically, a variety of prognosis scores have been developed as tools to stratify patient risk. Such scores output a single numeric number, which is ideally suited for ranking, triaging, and prioritizing care [5, 6, 7]. Naturally, metrics such as C-statistic [8] are appropriate to evaluate ranking tools, and log-ℓ_1 loss [7], and mean-squared-error [9] for point predictions. However, accurate and confident prognosis of a particular patient's future outcomes (as opposed to relative risk of this patient against others in a group) requires more nuanced forecasting of calibrated patient specific survival curves [10, 11, 12]. For example, when making a clinical decision of whether a given patient is at risk of a heart attack in the next 3 months vs the next 12 months, global metrics such as concordance are irrelevant, and sharp and calibrated predictions are needed instead [13, 14, 15].

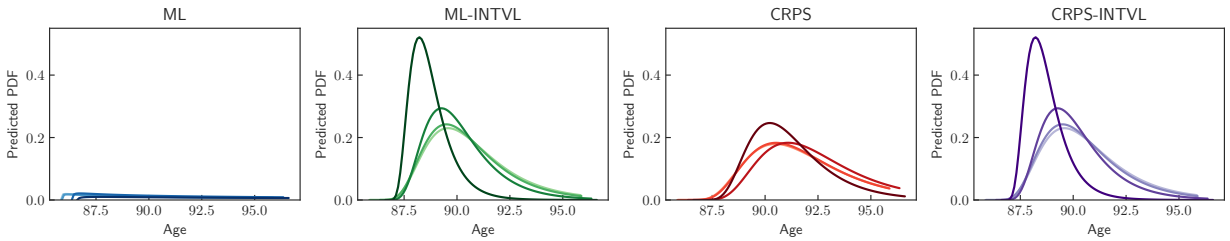


Figure 1: Example of a patient's predicted distributions for age of mortality under different objectives (losses). Repeated interactions are indicated by darker color. Our proposed techniques (b) and (d) improve sharpness of predicted distributions, while maintaining calibration from the original MLE methods (a) and (c). Graphs (a) and (b) compare MLE vs. CRPS in the right-censored setting and (c) vs. (d) in interval-censored setting (sharper the curves the better). The densities of (a) show that predicted variance from a model trained with MLE can be unacceptably high, especially in the heavily censored setting.

Calibration means that predicted probabilities over events match real-world frequencies of their occurrence. For example, among all the days that had a rain forecast of 80%, it should actually rain approximately 8 of 10 time. Probabilistic forecasts are accurate if they are well calibrated. However, calibration alone is not sufficient. Hypothetically, a model could always output the marginal distribution over outcomes as its prediction, making it well calibrated, but with little practical use. This is the notion captured by the paradigm of maximizing sharpness (i.e, concentration of probability) subject to calibration, which is widely and successfully adopted in meteorology [16]. The intuition behind this paradigm is that (i) uncalibrated predictions regardless of sharpness are wrong and useless, (ii) calibrated but non-sharp predictions are correct but less useful, and (iii) calibrated and sharp distributions are most useful (the sharper the better) [16]. For example, suppose a doctor is presented with predictions from multiple models, such as those shown in Figure 1. Although they all come from well calibrated models, it is clear that predictions from the model behind sharper distributions is more useful. Lack of sharpness is commonly encountered when training survival prediction models with MLE in the datasets that observe heavy censoring. Thus, naively using MLE based survival prediction models on large scale real-world data (where censored data is generally the majority) can be challenging.

The field of meteorology has improved the sharpness of probabilistic forecasts by abandoning the logarithmic scoring rule, and instead using the Continuous Ranked Probability Score (CRPS). The CRPS is a robust scoring rule which is not swayed by outliers as heavily as MLE. However CRPS is not as simple (neither analytically nor numerically) as MLE. Though CRPS has been used in the regression context to train models [16, 17, 18], it does not handle censored observations, which is crucial to building survival prediction models.

Summary of contributions:

(i) We introduce Survival-CRPS, a generalization of CRPS to handle right and interval-censored data (Section 2.1).

(ii) We propose a new evaluation metric, Survival-AUPRC, a generalization of the Area Under the Precision-Recall Curve that holistically measures sharpness and calibration, handling right-censored and interval-censored outcomes (Section 2.3).

(iii) We demonstrate the benefits of Survival-CRPS over MLE by producing sharper, and calibrated survival predictions of patient mortality on two large-scale EHR data sets (Section 3).

(iv) We provide practical recommendations and choices for implementing models with Survival-CRPS on large scale data (Section 2.4).

Though we frequently use the healthcare setting to describe ideas, all the concepts we present in this paper are completely general and apply to any right or interval censored survival prediction problem.

2 Countdown Regression

We consider a dataset of time-to-event records $\left\{x^{(i)}, y^{(i)}, c^{(i)}, \mathcal{T}^{(i)}\right\}$, where $x^{(i)} \in \mathbb{R}^d$ denotes a set of features, $y^{(i)} \in \mathbb{R}_+$ denotes time to event or censorship, $c^{(i)} \in \{0, 1\}$ is a censoring indicator where $c^{(i)} = 0$ means time to event is $y^{(i)}$, and $c^{(i)} = 1$ means time to event is at least $y^{(i)}$, and $\mathcal{T}^{(i)}$ denotes time by which the event must have occurred, where $\mathcal{T}^{(i)} = \infty$ in the right-censored setting and $\mathcal{T}^{(i)} \in \mathbb{R}_+$ in the interval-censored setting. We omit superscripts i for succinctness where possible in this section.

Parametric survival prediction methods model the time to an event of interest with a family of probability distributions, indexed by the distribution parameters. The survival function, denoted $S(t) : [0, \infty) \to [0, 1]$, is a monotonically decreasing function over the positive reals with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. The survival function represents the probability of an event of interest not occurring up to a given time. Every survival function has a corresponding cumulative density function (CDF), denoted $F(t) = 1 - S(t)$, and probability density function (PDF), denoted $f(t) = \frac{d}{dt}F(t)$. The choice of the family of probability distributions implies assumptions made about the nature of the data generating process.

2.1 SURVIVAL-CRPS: PROPER SCORING RULE OBJECTIVE

A scoring rule is a measure of the quality of a probabilistic forecast. A forecast over a continuous outcome is a probability density function over all possible outcomes, $\hat{f}$ with corresponding cumulative density function $\hat{F}$. In reality, we observe some actual outcome, y. A scoring rule $\mathcal{S}$ takes a predicted distribution and an actual outcome, and returns a loss $\mathcal{S}(\hat{F}, y)$. It is considered a *proper scoring rule* if for all possible distributions G,

$$\mathbb{E}_{y\sim\hat{F}}[\mathcal{S}(\hat{F}, y)] \leq \mathbb{E}_{y\sim\hat{F}}[\mathcal{S}(G, y)],$$

and *strictly* proper when equality holds if and only if $\hat{F} = G$ [16]. A proper scoring rule is one in which the expected score is minimized by the distribution with respect to which the expectation is taken. Intuitively, it encourages a model for being honest by predicting what it actually believes [19]. When a proper scoring rule is employed as a loss function, it naturally rewards the model for outputting calibrated probabilities [16].

There are many commonly used proper scoring rules. Perhaps the most widely used is the logarithmic scoring rule, equivalent to the MLE objective:

$$\mathcal{S}_{\text{MLE}}(\hat{F}, y) = -\log \hat{f}(y).$$

In the setting with censored data, we maximize the density for observed outcomes, and tail or interval mass for censored outcomes, and this is a proper scoring rule [20].

$$\begin{aligned}\mathcal{S}_{\text{MLE-RIGHT}}(\hat{F}, (y, c)) &= -\log\big((1-c)\hat{f}(y) \\ &\quad + c(1 - \hat{F}(y))\big) \\ \mathcal{S}_{\text{MLE-INTVL}}(\hat{F}, (y, c, \mathcal{T})) &= -\log\big((1-c)\hat{f}(y) \\ &\quad + c(\hat{F}(\mathcal{T}) - \hat{F}(y))\big)\end{aligned}$$

However, the logarithmic scoring rule is asymmetric, and harshly penalizes predictions that are wrong yet confident. Specifically, when the true data generating process is heavier tailed than the assumed data generating process, the training process becomes sensitive to outliers and yields more conservative (that is, less sharp) predictions as a result [4].

Another proper scoring rule for forecasts over continuous outcomes is the CRPS [21], defined as

$$\begin{aligned}\mathcal{S}_{\text{CRPS}}(\hat{F}, y) &= \int_{-\infty}^{\infty} \Big(\hat{F}(z) - \mathbb{1}\{z \geq y\}\Big)^2 dz \\ &= \int_{-\infty}^{y} \hat{F}(z)^2 dz + \int_{y}^{\infty} (1 - \hat{F}(z))^2 dz.\end{aligned}$$

The CRPS has been used in regression as an objective function that yields sharper predicted distributions compared to MLE, while maintaining calibration [16]. Intuition for the CRPS is better understood by analyzing the latter expression and noting that the two integral terms correspond to the two shaded regions in Figure 2a. The CRPS score is completely reduced to zero when the predicted distribution places all the mass on the point of true outcome, or equivalently, when the shaded region completely vanishes.

In the context of time to event predictions we propose the *Survival-CRPS* which accounts for the possibility of right-censored or interval-censored data:

$$\begin{aligned}\mathcal{S}_{\text{CRPS-RIGHT}}(\hat{F}, (y, c)) &= \int_{0}^{y} \hat{F}(z)^2 dz \\ &\quad + (1-c)\int_{y}^{\infty} (1 - \hat{F}(z))^2 dz, \\ \mathcal{S}_{\text{CRPS-INTVL}}(\hat{F}, (y, c, \mathcal{T})) &= \int_{0}^{y} \hat{F}(z)^2 dz \\ &\quad + (1-c)\int_{y}^{\mathcal{T}} (1 - \hat{F}(z))^2 dz \\ &\quad + \int_{\mathcal{T}}^{\infty} (1 - \hat{F}(z))^2 dz.\end{aligned}$$

Note that when $c = 0$, both of the above expressions are equivalent to the original CRPS. Again, the intuition behind the Survival-CRPS is better understood by mapping each of the integral terms to the corresponding shaded region in Figure 2b and Figure 2c. The Survival-CRPS behaves like the original CRPS when the time of event is uncensored. For censored outcomes, it penalizes the predicted mass that occurs before the time of censoring and, if interval censored, also the mass after time by which the event must have occurred.

Both variants of the Survival-CRPS are proper scoring rules. They are special cases of the threshold weighted CRPS [22], where the weighting function is an indicator over the uncensored regions.

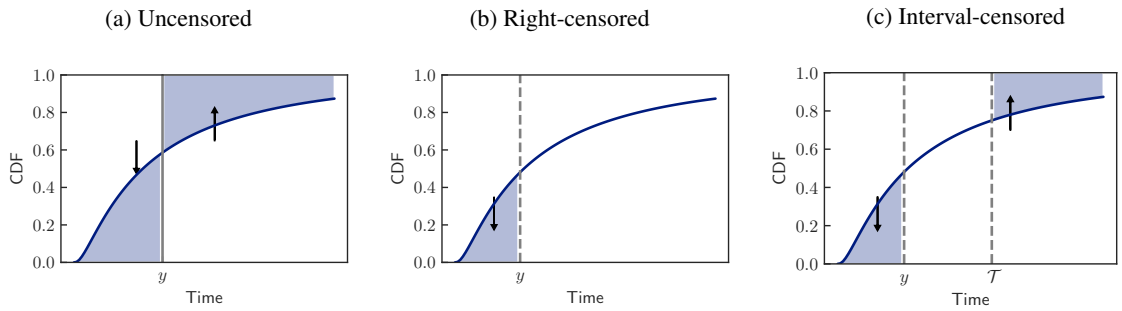


Figure 2: Graphical intuition for the Survival-CRPS scoring rule. For uncensored observations, we minimize mass before and after the observed time of event. For right-censored observations, we minimize mass before observed time of censoring. For interval-censored observations, we minimize mass before observed time of censoring, and mass after the time by which event must have occurred.

2.2 Evaluation by Sharpness subject to Calibration

Calibration assesses how well forecasted event probabilities match up to observed event probabilities. It is crucial in development of useful predictive models, especially for clinical decision-making. In binary prediction tasks without censoring, the Hosmer-Lemeshow test statistic [23] is commonly used to assess goodness-of-fit by comparing observed versus predicted event probabilities at quantiles of predicted probabilities. Extensions to account for censoring have been proposed [24, 25, 26], but these methods apply only to binary predictions for a particular time frame (for example, 1-year risks of mortality).

There is no widely accepted method for evaluating how well calibrated a set of entire prediction distributions is in the time to event setting. D-calibration has been recently proposed as a method for holistic evaluation [27], but relies on handling censored observations by assuming the true times to event are uniformly distributed past the times of censoring in the predicted distributions. When censored observations far outnumber the uncensored observations, this can lead to overly optimistic assessments of calibration. Another option is to evaluate observed event times on the cumulative density scale of predicted distributions, using a Kaplan-Meier estimate to account for censoring [1]. Again, this method has limitations in the heavily censored setting, as the quantiles in the tail of predicted cumulative densities have few uncensored observations, and will rarely yield well calibrated values.

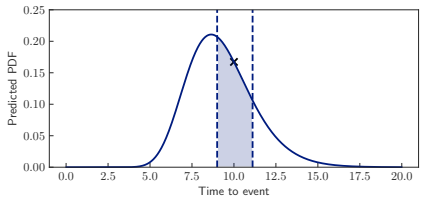
We instead employ the following method to measure calibration. We compare predicted cumulative densities against observed event frequencies, evaluated at quantiles of predicted cumulative density. Right-censored observations are removed from consideration in quantiles that correspond to times after their points of censoring. Interval-censored observations are similarly removed from consideration in quantiles that correspond to times after censoring, but are additionally re-introduced in quantiles that correspond to times past the time by which the event must have occurred (in the mortality prediction task, this corresponds to 120 years of age). In this work we assess resulting calibration curves qualitatively by graphing them, and quantitatively by comparing the slopes of the corresponding lines of best fit (ideally 1).

Subject to calibration, we strive for prediction distributions that are *sharp* (i.e, concentrated). There are several metrics that could be used for measuring sharpness, such as variance or entropy. In the context of time to event predictions, holding two distributions with vastly different means to the same standard of variance or entropy would be unfair (for example, we would want lower variance for a prediction distribution with a mean of a day, compared to a mean of a year). Instead, we use the coefficient of variation (CoV) as a reasonable measure of sharpness. The CoV is defined as the ratio of one standard deviation to the mean, $\text{CoV}(\hat{F}) = \frac{\sqrt{\text{Var}[\hat{F}]}}{\mathbb{E}[\hat{F}]}$.

2.3 Survival-AUPRC: Holistic Time to Event Metric

Since sharpness is only a function of the predicted distributions, a measure of sharpness is only meaningful if the model is sufficiently calibrated. We now propose a metric that measures how concentrated the mass of the prediction distribution is around the true outcome, robust to miscalibration. The idea is similar to the area under a precision-recall curve, except here it is with respect to only one predicted distribution and one outcome. We first consider the uncensored case. As an analog to precision, we consider intervals relative to the true time of event, defined by ratios. For example, a region of precision

(a) Analog of recall for precision of 0.9.



(b) Corresponding precision-recall plot.

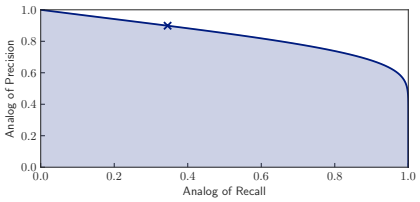


Figure 3: Graphical intuition for the Survival-AUPRC metric for uncensored observations. We compute the mass that lies within each interval of precision around the true outcome to compute the precision-recall curve.

0.9 around an event that occurs at time y is the interval $[0.9y, y/0.9]$. Corresponding to this region of precision, the analogy to recall is the mass assigned by the predicted distribution over this interval, $\hat{F}(y/0.9) - \hat{F}(0.9y)$. By exploring the full range of precision from 0 to 1, we obtain the *Survival Precision Recall Curve*. The area under this curve measures how quickly predicted mass concentrates around the true outcome as we expand the precision window.

$$\text{Surv-AUPRC}_{\text{UNCENS}}(\hat{F}, y) = \int_0^1 (\hat{F}(y/t) - \hat{F}(yt))dt$$

The highest possible score is 1, when the predicted distribution is a Dirac δ function centered over the time of outcome. The lowest possible score is 0, when the predicted distribution is infinitely dispersed. The mean of all Survival-AUPRC scores across examples provides an overall measure of the quality of the predictions.

The aforementioned metric only applies when the event outcome is uncensored. In the case of censored observations, we use the same analogy but with the right end of precision intervals defined with respect to the time by which the event must have occurred in the interval-censored case, or infinity in the right-censored case.

$$\text{Surv-AUPRC}_{\text{RIGHT}}(\hat{F}, y) = \int_0^1 (1 - \hat{F}(yt))dt$$

$$\text{Surv-AUPRC}_{\text{INTVL}}(\hat{F}, y, \mathcal{T}) = \int_0^1 (\hat{F}(\mathcal{T}/t) - \hat{F}(yt))dt$$

2.4 Implementation tips and choices

Common parametric distributions over time to event used in traditional survival analysis models include the Weibull, Log-Normal, Log-Logistic, and Gamma. In order to be sufficiently expressive in model space, we seek distributions with at least two parameters. We recommend the Log-Normal distribution because other distributions either involve the Gamma function in their density, or involve the pattern $(y/p_1)^{p_2}$, where p_1 and p_2 are parameters output from the neural network. We found these patterns to be highly sensitive to the inputs and to suffer from numerical instability issues.

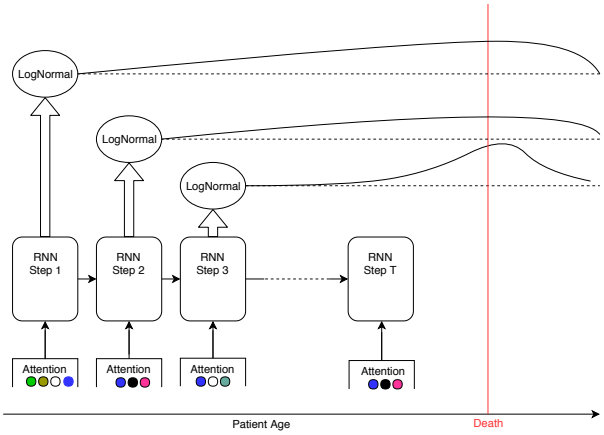
For the Log-Normal distribution, a closed form expression for the CRPS is well-known [28]. However, a closed form expression for the Survival-CRPS does not exist. We perform a change of variable to express the integral terms as finite integrals, and numerically approximate with the trapezoid rule. When training, we then back-propagate through the trapezoidal approximation. Details are given in Appendix B and C. Separately, we note that the approximation formulas are themselves proper scoring rules, as they are just weighted sums of Brier scores. Closed form expressions for the log-normal Survival-AUPRC are also given in Appendix D, E, and F.

3 Experiments

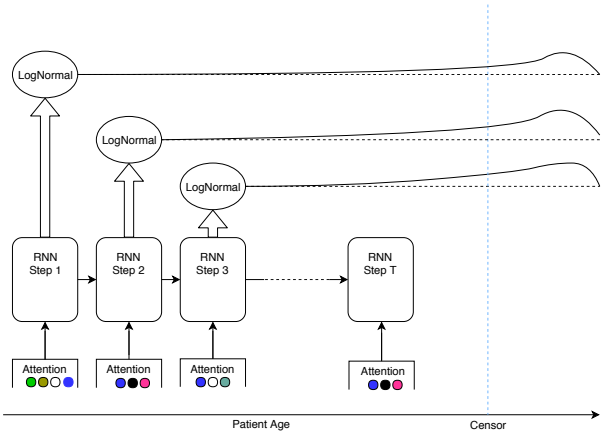
We assume a dataset of longitudinal records indexed by i, $\left(\{(x_t^{(i)}, a_t^{(i)})\}_{t=1}^{T^{(i)}}, d^{(i)}, c^{(i)}\right)$, where $t \in \{1 \dots T^{(i)}\}$ denotes the interaction number of this patient with the health record, $x_t^{(i)} \in \mathbb{R}^D$ is the set of features corresponding to the t-th interaction, $a_t^{(i)} \in \mathbb{R}_+$ is age at time t, $d^{(i)} \in \mathbb{R}_+$ is the age of death or age of last known (alive) encounter, and $c^{(i)} \in \{0, 1\}$ is a censoring indicator where $c^{(i)} = 0$ means the age of death is $d^{(i)}$, and $c^{(i)} = 1$ means the age of death is at least $d^{(i)}$. For each $x_t^{(i)}$ we define the quantity $y_t^{(i)} = d^{(i)} - a_t^{(i)}$ which represents the corresponding time to event or time to censoring. For interval censoring we assume a maximum lifespan of 120 years.

We run experiments for the mortality prediction task to evaluate four different training objectives: Maximum Likelihood $\mathcal{S}_{\text{MLE-RIGHT}}$ and $\mathcal{S}_{\text{MLE-INTVL}}$, and our Survival-CRPS based loss $\mathcal{S}_{\text{CRPS-RIGHT}}$ and $\mathcal{S}_{\text{CRPS-INTVL}}$.

(a) Dead (uncensored) patients.



(b) Alive (right-censored) patients.



(c) Alive (interval-censored) patients.

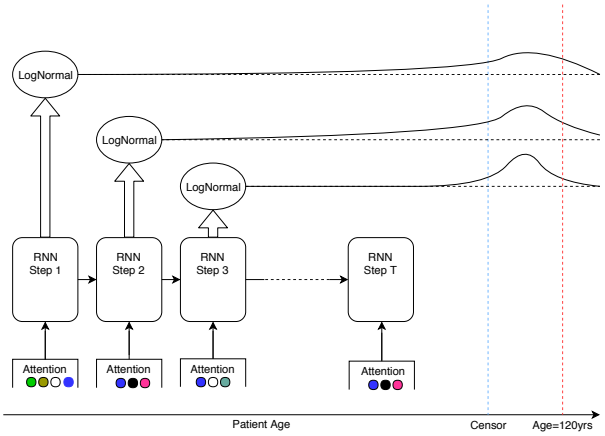


Figure 4: At each time step, we predict parameters μ, σ^2 of a log-normal distribution, minimizing a proper scoring rule. While (a) shows how to handle dead (uncensored) patients, (b) and (c) compare right and interval censorship. Leveraging knowledge about the world (e.g. mortality occurs by 120 yrs), (c) shows that interval censorship produces a sharper distribution.

3.1 RNN WITH STARR DATA WAREHOUSE

We use electronic health records from the STARR Data Warehouse (previously known as STRIDE) for training and evaluation [29]. The Warehouse contains de-identified data for over 3 million patients (about 2.6% having a recorded date of death), spanning approximately 27 years. The data we use on patients include diagnostic codes, medication order codes, lab test order codes, encounter type codes, and demographics (age and gender).

The structure of this dataset motivates our model selection; because many patients in this dataset have records spanning multiple days and visits, we use a recurrent neural network (RNN) on this particular task. We assign timesteps to each day that a patient has recorded data. The set of 3 million patients, correspond to 51 million overall timesteps, and was randomly split in the ratio 8:1:1 into train, validation and test splits. Formally, our model RNN is parameterized by θ, denoted RNN_θ, that takes as input a sequence of features to predict parameters of a parametric probability distribution $\hat{F}$ over time to death at each timestep (Figure 4). The network depends only on data from the current and previous timesteps, and not the future. The approach here is similar to the recently proposed Weibull Time to Event RNN [30], though we generalize to any choice of noise distribution. The distributions that are output in each timestep are used to construct an overall loss,

$$\mathcal{L}_{\text{RIGHT}} = \sum_{i=1}^{N} \sum_{t=1}^{T^{(i)}} \mathcal{S}_{\text{RIGHT}} \left(\hat{F}_{\text{RNN}_\theta\left\{x_{1:t}^{(i)}\right\}}, \left(y_t^{(i)}, c^{(i)}\right) \right)$$

$$\mathcal{L}_{\text{INTVL}} = \sum_{i=1}^{N} \sum_{t=1}^{T^{(i)}} \mathcal{S}_{\text{INTVL}} \left(\hat{F}_{\text{RNN}_\theta\left\{x_{1:t}^{(i)}\right\}}, \left(y_t^{(i)}, c^{(i)}, \mathcal{T}_t^{(i)}\right) \right),$$

where N is the total number of patients in the training set, $T^{(i)}$ is the sequence length for patient i, and $\hat{F}_{\text{RNN}_\theta}$ denotes the distribution parameterized by the output of the RNN.

We leverage a combination of real-valued (e.g. age of patient) and discrete (e.g., ICD codes) data. Discrete data is embedded into a trainable real-valued 126-dimensional vector space. The vectors of ICD codes at a single timestep are combined into a weighted mean by a soft self-attention mechanism. All real-valued inputs are appended to the averaged embedding vector. We also provide the real-valued features to every layer by appending them to the output of previous layer. The input vector feeds into a fully connected layer, followed by multiple recurrent layers. We use the Swish activation function [31] and layer normalization [32] at every layer. Recurrent layers are defined using GRU units [33] with layer normalization inside. After the set of recurrent layers, the network has multiple branches, one per parameter of the survival distribution (for the lognormal, μ and σ^2). The final layer in each branch has scalar output, optionally enforced positive with the softplus function, $\text{Softplus}(z) = \log(1+\exp(z))$. We use Bernoulli dropout [34] at all fully connected layers, and Variational RNN dropout [35] in the recurrent layers, with a dropout probability of 0.5. Optimization is performed using the Adam optimizer [36], with a fixed learning rate of 1e-3.

3.2 FCN WITH MIMIC-III

We leverage the publicly available MIMIC-III dataset [37]. This dataset differed from the STARR EHRs in that patients were often admitted only once and whose visits did not exceed a day. To match this structure of this dataset, we build a 4 layer feed forward neural network that takes in 51015 hospital admissions in the dataset (70.1% censored) and makes predictions at the time of discharge. We removed admissions where the patient's age was obfuscated or where the patient's discharge time occurred after their recorded date of death. As features, we used demographics (age as real-valued, and gender as binary) and embedded diagnostic codes into a 128-dimensional space.

3.3 RESULTS

The results are presented in Table 1. Both the coefficient of variation and the Survival-AUPRC metrics suggest that the Survival-CRPS with interval censoring yields the sharpest prediction distributions. Inspecting the probability past 120 years of age for the STARR dataset shows that a naively trained prediction model with MLE can assign more than 75% of the mass to unreasonable regions, which is highly undesirable for the purpose of prediction. We note that this behavior is largely due to low prevalence of uncensored examples, which is typical in real-world EHR data sets. As a result, the loss for the censored examples, which can be minimized by pushing mass as far away to the right as possible, dominates the small number of uncensored examples. While the benefits of Survival-CRPS are most pronounced in low prevalence datasets (STARR), they show benefit with moderate prevalence as well (MIMIC-III).

Since we employ proper scoring rules, the predictions tend to be well calibrated (Figure 5). By predicting an entire distribution over time to an indivdual's mortality, the same model can be used to make classification predictions at various time points, highlighting the flexibility of our approach. When evaluated at 6 month, 1 year, and 5 year probabilistic predictions of mortality, our model remains well-calibrated with high discriminative ability (Appendix G, Figure 6).

Source code of our implementation is published [1].

4 RELATED WORK

Recent work has demonstrated potential to significantly improve patient care by making predictions with deep learning models on EHR data [13, 14], but these works have been limited to treating the task as binary classification over a fixed time frame. Predicting survival curves instead of dichotomous outcomes has been explored [7, 12], but these works predict over a discrete set of times. Work in [38] also predicts full survival curves specific to a patient, but the use of Gaussian Processes makes it difficult to scale to datasets with millions of patients. Deep survival analysis [11] has been proposed, but is limited to a fixed shape Weibull (bypassing the concerns we raised about stability, but limited in expressivity). The work by [39] is similar to ours in terms of using log-normal noise distribution, but is limited to MLE training. DeepSurv [9] uses a Cox proportional hazards model, which similarly makes a set of inflexible assumptions. The WTTE-RNN [30] model is also limited to a Weibull distribution. All aforementioned models have only been optimized by MLE, instead of more robust proper scoring rules. The CRPS scoring rule has been used with neural networks in [40]. The work in [41] predicts survival curves (both non-parametric, and flexible flow based parametric curves) while also handling missing covariates. Another recent work [42] uses adversarial training for survival prediction. It has been shown that modern neural networks can be miscalibrated, and the work by [43] and [44] suggest ways to improve calibration.

[1] http://github.com/stanfordmlgroup/cdr-mimic

Table 1: Metrics measuring sharpness and calibration for models trained on the right-censored and interval-censored variants of the Maximum Likelihood and Survival-CRPS objectives. Confidence intervals are generated by bootstrap resampling of the test set.

STARR Dataset (97.4% censoring)

Metric	MLE-RIGHT	MLE-INTVL	CRPS-RIGHT	CRPS-INTVL
Calibration slope	1.125 ± 3e-4	1.139 ± 3e-4	1.003 ± 3e-4	0.959 ± 5e-4
Mean coefficient of variation	18.42 ± 5e-3	0.911 ± 4e-4	0.332 ± 1e-4	**0.301 ± 1e-4**
Mean prob of survival to age 120 yrs	0.754 ± 2e-5	0.045 ± 3e-5	0.015 ± 3e-5	**0.005 ± 1e-6**
Dead: mean Surv-AUPRC (uncen)	0.233 ± 2e-4	0.319 ± 3e-4	0.343 ± 4e-4	**0.366 ± 4e-4**
Alive: mean Surv-AUPRC (intvl-cen)	0.407 ± 6e-5	0.963 ± 2e-5	**0.977 ± 3e-5**	0.976 ± 3e-5

MIMIC-III Dataset (70.1% censoring)

Metric	MLE-RIGHT	MLE-INTVL	CRPS-RIGHT	CRPS-INTVL
Calibration slope	0.945 ± 9e-3	0.933 ± 1e-2	0.951 ± 1e-2	0.938 ± 1e-2
Mean coefficient of variation	2.218 ± 0.011	1.763 ± 0.006	1.797 ± 0.014	**1.647 ± 0.012**
Mean prob of survival to age 120 yrs	0.012 ± 4e-4	0.007 ± 2e-4	0.001 ± 2e-4	**0.001 ± 3e-5**
Dead: mean Surv-AUPRC (uncen)	0.329 ± 2e-3	0.338 ± 4e-3	0.342 ± 3e-3	**0.348 ± 4e-3**
Alive: mean Surv-AUPRC (intvl-cen)	0.993 ± 2e-4	0.999 ± 6e-5	1.000 ± 2e-4	**1.000 ± 1e-5**

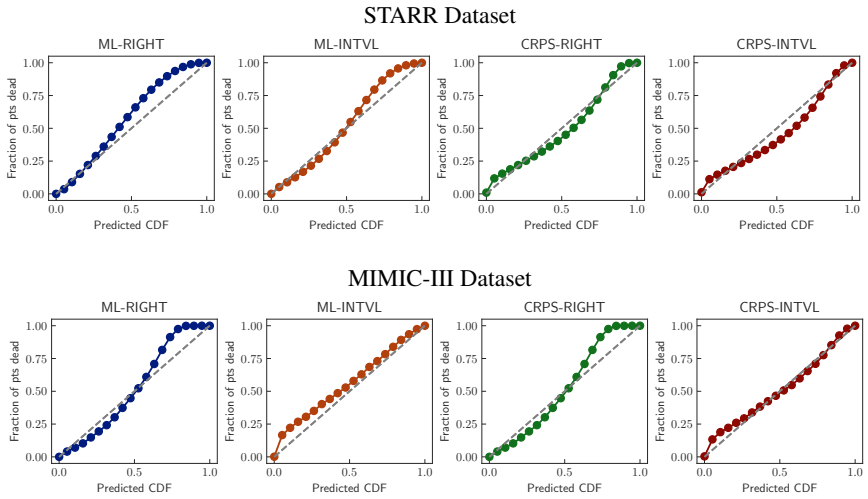


Figure 5: Calibration plots for each of the models. We compare predicted cumulative densities against observed event frequencies, evaluated at quantiles of predicted cumulative density. Right-censored observations are removed from consideration in quantiles past times of censoring, interval-censored observations are additionally re-introduced in quantiles corresponding to times past 120 years.

5 Conclusion

Better survival prediction models can be built by exploring objectives beyond MLE and evaluation metrics that assess the holistic quality of predicted distributions, instead of point estimates. We introduced the Survival-CRPS objective, motivated by the fact that the CRPS scoring rule is known to yield sharp prediction distributions while maintaining calibration. There are perhaps others scoring rules that work better, leaving avenues for future work. To evaluate, we introduced the Survival-AUPRC metric, which captures the degree to which a prediction distribution concentrates around the observed time of event. We demonstrate large-scale survival prediction by using a deep models employing a log-normal parameterization. The impact of having meaningfully accurate survival models is tremendous, especially in healthcare. We hope our work will be useful to those looking to build and deploy such models.

Acknowledgments

We thank Baran Sandor, Sebastian Lerch, Alejandro Schuler, Jeremy Irvin, and Russell Greiner for valuable feedback.

References

[1] Frank E. Harrell, Jr. *Regression Modeling Strategies*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Cox D. R. Regression models and life tables. *Journal of the Royal Statistic Society*, B(34):187–202, 1972.

[3] Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. In *Experimental Economics*, pages 43–61, 1998.

[4] Manuel Gebetsberger, Jakob W. Messner, Georg J. Mayr, and Achim Zeileis. Estimation Methods for Nonhomogeneous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood. *Monthly Weather Review*, 146(12):4323–4338, October 2018.

[5] Francis Lau, G. Michael Downing, Mary Lesperance, Jack Shaw, and Craig Kuziemsky. Use of palliative performance scale in end-of-life prognostication. *Journal of Palliative Medicine*, 9(5):1066–1075, 10 2006.

[6] Magnolia Cardona-Morrell and Ken Hillman. Development of a tool for defining and identifying the dying patient in hospital: Criteria for screening and triaging to appropriate alternative care (cristal). *BMJ supportive and palliative care*, 5(1):78—90, March 2015.

[7] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1845–1853. Curran Associates, Inc., 2011.

[8] Hajime Uno, Tianxi Cai, Lu Tian, and L. J. Wei. Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

[9] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1), December 2018. arXiv: 1606.00931.

[10] David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*, 129(25 suppl 2):S49–S73, June 2014.

[11] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep Survival Analysis. *arXiv:1608.02158 [cs, stat]*, August 2016. arXiv: 1608.02158.

[12] Changhee Lee, William Zame, and Jinsung Yoon. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. *AAAI*, page 8, 2018.

[13] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. Improving palliative care with deep learning. pages 311–316. IEEE, November 2017.

[14] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Peter J. Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin E. Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *arXiv:1801.07860 [cs]*, January 2018. arXiv: 1801.07860.

[15] Eli Sherman, Hitinder S. Gurm, Ulysses J. Balis, Scott R. Owens, and Jenna Wiens. Leveraging Clinical Time-Series Data for Prediction: A Cautionary Tale. In *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, November 4-8, 2017*, 2017.

[16] Tilmann Gneiting and Matthias Katzfuss. Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

[17] Seyedeh Atefeh Mohammadi, Morteza Rahmani, and Majid Azadi. Meta-heuristic CRPS minimization for the calibration of short-range probabilistic forecasts. *Meteorology and Atmospheric Physics; Wien*, 128(4):429–440, August 2016.

[18] Seyedeh Atefeh Mohammadi, Morteza Rahmani, and Majid Azadi. Optimization of continuous ranked probability score using PSO, 2015.

[19] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[20] A. Philip Dawid and Monica Musio. Theory and Applications of Proper Scoring Rules. *METRON*, 72(2):169–183, August 2014. arXiv: 1401.0398.

[21] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(2):243–268, 4 2007.

[22] Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

[23] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition*. Wiley Blackwell, 10 2011.

[24] Jon Ketil Grønnesby and Ørnulf Borgan. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis*, 2(4):315–328, Dec 1996.

[25] R.B. D'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models: Discrimination and calibration measures. In *Advances in Survival Analysis*, volume 23 of *Handbook of Statistics*, pages 1 – 25. Elsevier, 2003.

[26] Olga V. Demler, Nina P. Paynter, and Nancy R. Cook. Tests of Calibration and Goodness of Fit in the Survival Setting. *Statistics in medicine*, 34(10):1659–1680, May 2015.

[27] Axel Andres, Aldo Montano-Loza, Russell Greiner, Max Uhlich, Ping Jin, Bret Hoehn, David Bigam, James Andrew Mark Shapiro, and Norman Mark Kneteman. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PLOS ONE*, 13(3):e0193523, March 2018.

[28] Sándor Baran and Sebastian Lerch. Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141(691):2289–2299, mar 2015.

[29] Henry J Lowe, Todd A Ferris, Penni M Hernandez Nd, and Susan C Weber. STRIDE – An Integrated Standards-Based Translational Research Informatics Platform. *AMIA Annual Symposium Proceedings*, pages 391–395, 2009.

[30] Egil Martinsson. A model for sequential prediction of time-to-event in the case of discrete or continuous censored data, recurrent events or time-varying covariates. page 103, 2016.

[31] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017.

[32] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[33] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[34] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[35] Y. Gal and Z. Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *ArXiv e-prints*, December 2015.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[37] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035+, May 2016.

[38] Ahmed M. Alaa and Michaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2329–2337. Curran Associates, Inc., 2017.

[39] Yinchong Yang, Peter A. Fasching, and Volker Tresp. Modeling progression free survival in breast cancer with tensorized recurrent neural networks and accelerated failure time models. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 164–176, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

[40] Stephan Rasp and Sebastian Lerch. Neural networks for post-processing ensemble weather forecasts. abs/1805.09091, 2018.

[41] Xenia Miscouridou, Adler J. Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis : Nonparametrics and missingness. 2018.

[42] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence

Carin, and Ricardo Henao. Adversarial time-to-event modeling. In *ICML*, 2018.

[43] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[44] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *International Conference on Machine Learning*, pages 2796–2804, July 2018.