HHS Public Access

Author manuscript

IEEE Trans Neural Netw Learn Syst. Author manuscript; available in PMC 2022 April 29.

Published in final edited form as:

IEEE Trans Neural Netw Learn Syst. 2023 April ; 34(4): 1666–1680. doi:10.1109/TNNLS.2020.3029631.

Calibration and Uncertainty in Neural Time-to-Event Modeling

Paidamoyo Chapfuwa,

Chenyang Tao,

Lawrence Carin,

Ricardo Henao

Duke University

Abstract

Models for predicting the time of a future event are crucial for risk assessment, across a diverse range of applications. Existing time-to-event (survival) models have focused primarily on preserving pairwise ordering of estimated event times, or relative risk. Model calibration is relatively under explored, despite its critical importance in time-to-event applications. We present a survival function estimator for probabilistic predictions in time-to-event models, based on a neural network model for draws from the distribution of event times, without explicit assumptions on the form of the distribution. This is done like in adversarial learning, but we achieve learning without a discriminator or adversarial objective. The proposed estimator can be used in practice as a means of estimating and comparing conditional survival distributions, while accounting for the predictive uncertainty of probabilistic models. Extensive experiments show that the proposed model outperforms existing approaches, trained both with and without adversarial learning, in terms of both calibration and concentration of time-to-event distributions.

1 Introduction

Time-to-event studies aim to characterize the covariate effects on the time of a future event, while capitalizing on information from censored events when performing learning. Conventional nonparametric time-to-event (also called survival) models primarily involve methods that maximize the Concordance Index (C-Index) [22], a metric related to the receiver operating characteristic, that quantifies the degree to which estimated events result in pairwise orderings that are consistent with observed event times, *i.e.,* the ground truth. Consequently, any model that is able to estimate properly ordered but *proportional* event times can score high in terms of C-Index. A prominent example is the widely used Cox Proportional Hazards (CPH) model [9].

Predicting temporally accurate event times is important in a variety of applications, *e.g.,* risk profiling [23, 40], drug development [14], and prevention of online fraudulent activities [55]. Estimating temporally accurate event times typically involves the use of parametric Maximum Likelihood Estimation (MLE) approaches [30] or recently-developed nonparametric sampling based methods, e.g., via adversarial learning [7] or normalizing

paidamoyo.chapfuwa@duke.edu .

flows [36]. Further, given the critical time-sensitive nature of time-to-event modeling, it is highly desirable to design models that are not only temporally accurate but also produce *population-calibrated* and *uncertainty-aware* predictions.

Classical survival models include the CPH semiparametric model [9] that learns *relative* risk (proportional to time-to-event) as a function of covariates, and the Accelerated Failure Time (AFT) model [50], a parametric specification for temporally accurate event times that assumes covariates either accelerate or decelerate the progression of event time. AFT often assumes log-normal distributed event times, however, other likelihood functions have been considered, *e.g.*, exponential, Gamma, Weibull, *etc.* [4, 30]. These classical approaches assume a linear relationship between event times and covariates, which may be limiting for modern, large and highly heterogeneous datasets.

Time-to-event methods based on deep-learning are often direct extensions of classical models that aim to learn more flexible, non-linear mappings between event times and covariates. CPH-based deep learning methods [27, 56] have demonstrated improvements in C-Index relative to classical approaches, in some settings. Parametric extensions include the Deep Regularized Accelerated Failure Time (DRAFT) model [7], Deep Survival Analysis (DSA) [41] and the Survival Continuous Ranked Probability Score (S-CRPS) model [2]. Nonparametric extensions include Deep Adversarial Time-to-Event (DATE) [7], nonparametric DSA [36] and Gaussian-process-based models [1, 13, 34]. As an alternative to a strict time-to-event formulation, some approaches discretize event times and specify models that predict the probability of survival at discrete intervals [15, 35, 52].

Methods that produce uncertainty-aware predictions aim to estimate time-to-event *distributions,* rather than point estimates. Most approaches, parametric or not, result in either a parametric time-to-event distribution, *e.g.,* log-normal in AFT, DRAFT and S-CRPS and Weibull in DSA, or samples from an implicitly defined distribution, *e.g.,* DATE and nonparametric DSA. The latter uses normalizing flows. Importantly, uncertainty-aware predictions are only useful if the time-to-event distributions are concentrated, *i.e.,* their probability masses have coverage much smaller than the observed time range. This is key, because only in that case can uncertainty be leveraged effectively for ranking or prioritizing events/subjects. However, only a few approaches have considered the uncertainty of the predictions when assessing performance, namely, [7] via distribution coverage and [2] via coefficient-of-variation metrics.

Calibration, a descriptor of a predictive model that characterizes the statistical consistency of the predictions relative to the distribution of the observations on a population level, has been studied in forecasting [11], Bayesian analysis [10], and in machine learning, for classification [21] and regression [33] problems. Unfortunately, it is under-explored in time-to-event models. Exceptions include [3, 34, 49, 54] that use (time horizon) thresholded time-to-event Brier scores to asses calibration [6], and [2] that use calibration slope as a way to compare model performance. Note that although Brier scores are often used to assess calibration, most commonly in classification models, summaries of calibration curves such as the calibration slope are usually considered more informative [47].

We present an approach that *implicitly* defines time-to-event distributions conditioned on covariates via a neural network specification, from which we can synthesize temporally accurate, concentrated and calibrated time-to-event distributions. To this end, *i*) we present a reinterpretation of the Kaplan-Meier estimator for survival functions; *ii*) we extend it to estimate survival functions *conditional* on covariates; *iii*) we show that the new estimator can be used for visual and quantitative assessment of calibration; *iv*) we propose using it as an objective function in a neural-network-based nonparametric time-to-event model, to encourage calibrated predictions; *v*) we directly match the conditional survival function of the model to that of the ground truth without the need of adversarial techniques [18]; *vi*) we show that our *survival function matching* approach is related to earth mover's distance minimization; and *vii*) we present extensive quantitative and qualitative results, showing that our approach outperforms existing time-to-event models in terms of calibration, while being competitive in terms of C-Index and concentration (sharpness) of the predicted time-to-event distributions.

2 Background

Assume a time-to-event dataset, $\mathscr{D} = \{\boldsymbol{x}_n, t_n, y_n\}_{n=1}^{N}$, consisting of N observations (or subjects). For the n-th observation, we have d covariates, $\boldsymbol{x}_n = [x_{1n}, \ldots, x_{dn}] \in \mathbb{R}^d$, a time point, t_n, and a censoring indicator, $y_n \in \{0, 1\}$. When $y_n = 1$, t_n represents the *time-to-event* of interest, and when $y_n = 0$, t_n is the censoring time. Typically, events are *right* censored, meaning that given $y_n = 0$, **all we know about the** n-th observation is that we have not observed the event of interest up to time t_n. Though *left* and *interval* censored events are possible, these are far less common and are thus not usually considered in practice. Here we only consider right censoring, however, the proposed approach is general and can be readily extended using ideas from [2].

Time-to-event (or survival) models either characterize the conditional survival function $S(t|\boldsymbol{x})$, time density $f(t|\boldsymbol{x})$, or the hazards function $h(t|\boldsymbol{x})$, where the conditioning is on covariates $\boldsymbol{x}$. The survival function $S(t|\boldsymbol{x}) = P(\boldsymbol{\tau} > t|\boldsymbol{x})$, for $\boldsymbol{\tau} > 0$, which can also be written as the complement of the conditional cumulative density function, $F(t|\boldsymbol{x})$; hence, $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$ is a monotonically decreasing function of time. Learning the time-to-event conditional distribution, $f(t|\boldsymbol{x})$, can in principle yield both $S(t|\boldsymbol{x})$ and $h(t|\boldsymbol{x})$, provided $f(t|\boldsymbol{x}) = S(t|\boldsymbol{x})h(t|\boldsymbol{x})$. For some parametric choices of the conditional density, $f(t|\boldsymbol{x})$, the survival and hazards functions can be obtained in closed-form [30]. For instance, assuming the exponential density, $f(t|\boldsymbol{x}) = \lambda_{\boldsymbol{x}} \exp(-\lambda_{\boldsymbol{x}} t)$, yields $h(t|\boldsymbol{x}) = \lambda_{\boldsymbol{x}}$ and $S(t|\boldsymbol{x}) = \exp(-\lambda_{\boldsymbol{x}} t)$, where $\lambda_{\boldsymbol{x}}$ is a function of $\boldsymbol{x}$. See [4] for a few other examples. In practice, we seek to approximate the time density $f(t|\boldsymbol{x})$ with $q(t|\boldsymbol{x})$, a function parametrically or nonparametrically specified and learned from data, $\mathscr{D}$. The dataset $\mathscr{D}$ represents the *ground truth* or, conceptually, the empirical joint distribution $p(t, y, \boldsymbol{x})$ with marginals $p(t)$, $p(y)$ and $p(\boldsymbol{x})$, from which $p(t)$ is of most interest in our case, as described below.

The Kaplan-Meier Estimator

The standard Kaplan-Meier (KM) estimator [26] is a widely-used frequentist approach to estimate the (marginal) survival function, $S(t)$, using samples from $p(t)$, *i.e.,* the time-to-

event empirical distribution. Let $\mathcal{T} = \{t_i | t_i > t_{i-1} > \ldots > t_0\}$ be the set of distinct and ordered observed event times (censored and non-censored). The KM estimate for time t_i can be evaluated recursively as

$$\hat{S}_{\text{KM}}(t_i) = \left(1 - \frac{d_i}{n_i}\right)\hat{S}_{\text{KM}}(t_{i-1}), \tag{1}$$

where n_i is the number of subjects *at risk* at the beginning of follow-up interval $[t_i, t_{i+1})$, d_i is the number of non-censored events that occur within the same interval, $[t_i, t_{i+1})$, and $\hat{S}_{\text{KM}}(t_0) = 1$, indicating that at t_0 there are no observed events so $d_0 = 0$ and $n_0 = N$. It has been shown [24] that the KM estimator can be interpreted as a random process, where the number of events, d_i, within each discrete interval $[t_i, t_{i+1})$ can be modeled as a draw from a Binomial distribution $d_i \sim \text{Binomial}(n_i, \pi)$, with mean event rate π. Moreover, it has been proven that KM is a consistent estimator [38], *i.e.*, $\sqrt{N}(\hat{S}_{\text{KM}}(t) - S(t))$ converges to a Gaussian process [5], with zero mean and covariance function recursively approximated by Greenwood's formula [20].

Distribution-Based Kaplan-Meier Estimator

The standard KM estimator is a *population* statistic that approximates the marginal survival distribution $S(t)$. Consequently, KM does not explicitly accommodate the use of predictions, *i.e.*, individualized (subject-level) conditional survival functions. Considering that time-to-event methods are primarily tasked with individualized predictions of conditional time densities, $f(t|\boldsymbol{x})$, which can be then used to obtain conditional survival functions $S(t|\boldsymbol{x})$, below we present a modified KM estimator that accounts for individualized time-to-event predictions.

We first consider a KM estimator for *point estimates* of $S(t|\boldsymbol{x})$, directly formulated from the standard KM in (1). It is then extended to probabilistic, *distribution* estimates of $S(t|\boldsymbol{x})$. The point-estimate-based KM, denoted PKM, estimates the population survival function accounting for covariates using predictions $\hat{T}_n \sim g(\boldsymbol{x}_n)$, where $g(\boldsymbol{x}_n)$ is some predictive function, or a summary from a probabilistic estimate of the conditional density $f(t|\boldsymbol{x}_n)$, *e.g.,* $\hat{T}_n \sim g(q(t|\boldsymbol{x}_n))$ where $g(\cdot) = \text{mean}(\cdot)$ and $q(t|\boldsymbol{x}_n)$ is the approximated conditional learned from dataset $\mathcal{D}$. We then write

$$\hat{S}_{\text{PKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} \mathbb{I}\left(t_{i-1} \le \hat{T}_n < t_i\right)}{N - \sum_{n=1}^{N} \mathbb{I}\left(\hat{T}_n < t_{i-1}\right)}\right)\hat{S}_{\text{PKM}}(t_{i-1}), \tag{2}$$

where $\hat{S}_{\text{PKM}}(t_0) = 1$, $\mathbb{I}(a)$ is an indicator function such that $\mathbb{I}(a) = 1$ if a holds or $\mathbb{I}(a) = 0$ otherwise. It follows from (2) that $\hat{S}_{\text{PKM}}(t_i) = \hat{S}_{\text{KM}}(t_i)$, when $\hat{T}_n$ represents an observed (ground truth) time-to-event from $p(t)$.

To account for predictive uncertainly, *i.e.,* for probabilistic estimates $q(t|\boldsymbol{x}_n)$, we extend (2) to distribution-based Kaplan-Meier (DKM) estimator. Specifically, we write

$$\hat{S}_{\text{DKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} F_n(t_i|\boldsymbol{x}_n) - F_n(t_{i-1}|\boldsymbol{x}_n)}{N - \sum_{j=1} F_n(t_{i-1}|\boldsymbol{x}_n)}\right)\hat{S}_{\text{DKM}}(t_{t-1}), \quad (3)$$

where $F_n(t_i|\boldsymbol{x}_n)$ is the estimated cumulative density function for subject n conditioned on covariates $\boldsymbol{x}_n$ and evaluated at t_i. Note that $\hat{S}_{\text{DKM}}(t_i) = \mathbb{E}_{q(t|\boldsymbol{x}_1)\ldots q(t|\boldsymbol{x}_N)}[\hat{S}_{\text{PKM}}(t_i)]$ so (3) averages over (samples of) $q(t|\boldsymbol{x}_n)$ rather than being evaluated on summaries (e.g., averages) of $q(t|\boldsymbol{x}_n)$ as in (2). For probabilistic estimates $q(t|\boldsymbol{x}_n)$ of $f(t|\boldsymbol{x}_n)$, the estimator in (3) is attractive because it accounts for the predictive uncertainty of the model, thus on a population level, it comprehensively captures the uncertainty of the estimated conditional survival distribution.

Calibration in Time-to-Event Models

In the context of time-to-event modeling, calibration refers to the concept of obtaining a predictor of time-to-event (that may or may not be probabilistic) whose predictions match, on a population level, the survival distribution $S(t)$. Figure 1 shows estimated survival distributions on the support dataset (see Section 5 for details) for five different models (DATE, DRAFT, SFM, CPH and S-CRPS) using DKM in (3), as well as the ground truth (Empirical) using KM in (1). Error bars (shaded regions) are calculated using the exponential Greenwood's formula [24].

From Figure 1, we see that DKM in (3) can be used to visually assess the calibration of estimated event times from different models relative to the ground truth. Specifically, we see that one of the models, SFM (the proposed model) matches the ground truth (Empirical) substantially better than the alternatives (see Section 5 for details). Strikingly, the other three models underestimate survival almost everywhere. In the experiments, we will use KM and DKM to more directly visualize calibration, and summarize it in terms of *calibration slope*. Further, below we leverage DKM to encourage calibration during model training, i.e., that DKM for a given model that approximates $q(t, \boldsymbol{x}_n)$ matches as well as possible the true survival distribution estimated via KM.

3 Survival Function Matching

We propose a nonparametric model for survival-function matching. Specifically, we approximate the density $f(t|\boldsymbol{x})$ implicitly as $q(t|\boldsymbol{x})$ via deterministic function $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$ which we specify as a neural network parameterized by $\boldsymbol{\theta}$ and where ϵ is a source of stochasticity, distributed according to some simple distribution, *e.g.,* uniform or Gaussian. In this manner, we do not impose/assume an explicit form on $q(t|\boldsymbol{x})$, we only seek to efficiently synthesize samples from it. This type of model has been considered recently within an adversarial-learning setup [7], but in the proposed work adversarial learning is not required, thus simplifying learning. Further, [7] did not consider calibration.

Calibration objective

Assume as above that $\mathcal{T}$ is the set of distinct and ordered observed event times (censored or non-censored). To estimate the parameters of the model $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$ that generates time-to-event samples on a population level, we match synthesized samples to the empirical survival

function, $S(t)$, thus producing calibrated predictions. We propose optimizing the following objective

$$\ell_{\text{cal}}(\boldsymbol{\theta}; \mathscr{D}) = \frac{1}{|\mathscr{T}|} \sum_{t_i \in \mathscr{T}} \left\| \hat{S}_{\text{PKM}}^{p(t)}(t_i) - \hat{S}_{\text{PKM}}^{G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)}(t_i) \right\|_1, \tag{4}$$

where $|\mathscr{T}|$ is the cardinality of $\mathscr{T}$, and $\hat{S}_{\text{PKM}}^{p(t)}(t_i)$ and $\hat{S}_{\text{PKM}}^{G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)}(t_i)$ are obtained from $p(t)$ and samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$, respectively. This is connected to KM, because $\hat{S}_{\text{PKM}}^{p(t)}(t_i)$ are obtained from $p(t)$.

The objective in (4) seeks to obtain model parameters, $\boldsymbol{\theta}$, for which model and empirical survival functions match. Note that the objective accounts for both censoring and non-censored events. Provided that the conditional survival distribution $S(t|\boldsymbol{x}) = P(\tau > t|\boldsymbol{x})$for $\tau \geq 0$ is the complement of conditional cumulative density function $F(t|\boldsymbol{x})$, matching the conditional survival function also matches the conditional time density $f(t|\boldsymbol{x})$, *i.e.,* the time-to-event distribution.

Learning with (4) is challenging because $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathscr{D})$ is a discrete function, and thus backpropagation is difficult. Several techniques have been developed to efficiently obtain unbiased and low-variance gradients for backpropagation with discrete objectives or sampling distributions, thus alleviating some of its challenges. Such techniques include REINFORCE [51], reparameterization tricks [29, 42], and more recently RELAX [19], a technique that combines both REINFORCE and reparameterization tricks via a variance-reduction neural network.

To circumvent the challenges of optimizing over the discrete function in (2) and favor simplicity, we instead optimize over its expectation in (3), which is continuous. However, replacing (2) with (3) is not only inefficient, as it requires generating multiple samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$, but also challenging because $F(t|\boldsymbol{x})$, the conditional cumulative function for $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$ is not available in closed-form. Conveniently, we can replace the indicator functions $\mathbb{I}(a)$ in (2) with Heaviside step functions, $H(b) = \frac{1}{2}(\text{sign}(b) + 1)$, therefore obtaining a differentiable formulation:

$$\hat{S}_{\text{PKM}}(t_i) = \left(1 - \frac{\sum_{n:y_n=1} H\left(\hat{T}_n - t_{i-1}\right) - H\left(\hat{T}_n - t_i\right)}{N - \sum_{n=1}^{N} H\left(t_{i-1} - \hat{T}_n\right)}\right) \hat{S}_{\text{PKM}}(t_{i-1}) \,. \tag{5}$$

When evaluating the objective, $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathscr{D})$ in (4), $\hat{T}_n$ is either a sample from the model, $\hat{T}_n = G_{\theta}(\boldsymbol{x}_n, \epsilon)$, or an observed time $\hat{T}_n \sim p(t)$, for $\hat{S}_{\text{PKM}}^{G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)}(t_i)$ or $\hat{S}_{\text{PKM}}^{p(t)}(t_i)$, respectively.

Accuracy objective

The objective $\ell_{\text{cal}}(\boldsymbol{\theta}; \mathscr{D})$ in (4) optimizes over a population estimate that encourages calibration. However, calibration alone does not result in time-to-event samples from $G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)$ that are accurate or concentrated wrt the ground truth. This happens because, for a given problem, there exist many solutions that yield well-calibrated predictions that are not necessarily accurate, thus not practically useful. For instance, take the extreme case

for which a model learns to estimate $p(t)$ independent of (ignoring) the covariates, $\boldsymbol{x}$, thus effectively recovering the KM estimator in 1. So motivated, we also specify accuracy-enforcing objective functions for censored and non-censored observations by borrowing from the recently proposed DATE model [7]. Specifically, we split the dataset $\mathscr{D}$ into two disjoint sets $\mathscr{D}_c$ and $\mathscr{D}_{nc}$, for censored and non-censored observations, respectively, and let $(t, \boldsymbol{x}) \sim p_C$ and $(t, \boldsymbol{x}) \sim p_{nc}$ represent, respectively, empirical distributions for these sets. We write objective functions for $\mathscr{D}_c$ and $\mathscr{D}_{nc}$ as

$$\ell_{\text{acc}}(\boldsymbol{\theta}; \mathscr{D}_c, \mathscr{D}_{nc}) = \mathbb{E}_{(t,\boldsymbol{x}) \sim p_c, \epsilon \sim p_\epsilon}[\max(0, t - G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)] + \mathbb{E}_{(t,\boldsymbol{x}) \sim p_{nc}, \epsilon \sim p_\epsilon}[|t - G_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon)|], \tag{6}$$

where $\epsilon \sim p_\epsilon$ has a simple distribution (uniform or Gaussian), $\max(0, \cdot)$ in the first term encourages that time-to-event samples from the model, evaluated on censored observations $y_n = 0$, are larger than the censoring time. The second term, absolute error, encourages time-to-event samples to be accurate, *i.e.,* as close as possible to the ground truth, for non-censored (observed) observations.

Consolidated objective

The complete objective function for the proposed Survival Function Matching (SFM) model is $\ell(\boldsymbol{\theta}; \mathscr{D}) = \ell_{\text{cal}}(\boldsymbol{\theta}; \mathscr{D}) + \lambda \ell_{\text{acc}}(\boldsymbol{\theta}; \mathscr{D}_c, \mathscr{D}_{nc})$, where $\lambda > 0$ is a free parameter controlling the trade-off between the accuracy objective and the survival function matching objective in (4). In the experiments we let $\lambda = 1$, however, λ can be optimized by grid search if desired.

The complete objective is optimized using stochastic gradient descent on minibatches from $\mathscr{D}$. Note that $\ell_{\text{cal}}(\theta; \mathscr{D})$ is a population-level objective that may be affected by the minibatch size, however, empirically we did not observe substantial differences in the performance metrics when varying the minibatch size (see the Supplementary Material). We justify the model being insensitive to the minibatch size owing to the insight that learning with minibatches can be understood as encouraging the model to be calibrated for every minibatch, thus consequently also encouraging global calibration.

4 Related Work

Existing calibration literature in predictive models has primarily focused on recalibration techniques for predictions from classification [21] or regression models [33]. For classification tasks, the Brier score [6] is a commonly used proper score metric, quantifying the accuracy of probabilistic predictions, and thus it is often used to assess calibration. The Brier score has also been used to asses calibration in time-to-event models [3, 34, 49, 54], however, this score has to be evaluated at pre-specified (thresholded) time horizons. Alternatively, S-CRPS [2] considers the integral of the Brier score evaluated at all possible thresholds [17], which is a more principled and comprehensive approach than calibration at pre-specifying time horizon thresholds.

Please add a Supplementary Material subhead at the end of the manuscript with the following paragraph text: Supplementary material may be found at: https://github.com/paidamoyo/adversarial_time_to_event.

The approach presented here is inspired by [2]. They considered calibration slope as a metric for evaluating performance in time-to-event models. However, our formulation is very different from that of [2], in the sense that they encourage calibration by optimizing a proper score rule, the Continuous Ranked Probability Score (CRPS), whereas we tackle it directly as a survival-distribution-matching problem. In the experiments in Section 5, we show empirically that our more direct approach to calibration consistently outperforms CRPS.

Our work is related to other recently proposed time-to-event approaches, including Survival CRPS (S-CRPS) [2], that uses an AFT (log-normal distribution) specification; the *conditional-GAN* approach [DATE, 7]; the AFT-based DRAFT model [7]; the Weibull-based Deep Survival Analysis (DSA) [41]; nonparametric DSA based on *normalizing flows* [36]; and Gaussian-process-based models [1, 13, 34]. Interestingly, excluding approaches that address *thresholded* calibration with Brier-scores [3, 34, 49, 54], only S-CRPS [2] considers *global* calibration as a performance metric. All the others focus on accuracy-centric performance estimates, e.g., C-Index and relative absolute error.

Optimal mass transport approaches for distribution matching in machine learning tasks have received considerable attention recently [8, 45]. For one-dimensional problems, it has been shown that the characterization of the p-Wasserstein metric has a simple form [32] $W_p(P,Q) = \left(\int_0^1 \left|F^{-1}(z) - G^{-1}(z)\right|^p dz\right)^{1/p}$ where, $F(z)^{-1}$ and $G(z)^{-1}$, for $z \in (0, 1)$, are the *quantile functions* of $p(t)$ and $q(t)$, respectively, and $F(t)$ and $G(t)$, their corresponding cumulative density functions. Interestingly for $p=1$, $W_p(P, Q)$ is also known as the *Monge-Rubenstein metric* [48] or the *earth mover's distance* [44], and it is essentially the absolute difference between the quantile functions for $p(t)$ and $q(t)$. By contrast, the SFM objective in (4) is the absolute difference between the cumulative density functions for $p(t)$ and $q(t)$, provided that $F(t) = 1 - S(t)$. As a result, minimizing (4) and $W_p(P, Q)$ are closely related approaches to matching $p(t)$ and $q(t)$. However, the latter explicitly imposes survival-distribution matching, which we consider more appropriate considering the goal is to obtain calibrated predictions in the context of time-to-event modeling.

5 Experiments

We qualitatively and quantitatively compare the proposed approach, SFM, against DATE [7], DRAFT [7], and CPH [9] and S-CRPS [2]. Complete details of model architectures, optimization, validation and testing are in the Supplementary Material.

Datasets

We consider five diverse datasets: *i*) flchain: a public dataset investigating non-clonal serum immunoglobin free light chains effects on survival time [12]. *ii*) SUPPORT: a public dataset for a survival-time study of seriously-ill hospitalized adults [31]. *iii*) seer: a public dataset provided by the Surveillance, Epidemiology, and End Results (SEER) Program. We restrict the dataset to a 10-year follow-up breast cancer subcohort with three competing risks (breast cancer, cardiovascular and others). See [43] for preprocessing details. *iv)* EHR: a large study from the Duke University Health System centered around multiple inpatient visits

due to comorbidities in patients with Type-2 diabetes [7]. *v*) sleep: a subset of the Sleep Heart Health Study (SHHS) [39], a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing.

Table 1 presents summary statistics of the datasets, where d denotes the size of the *individual* covariate vector x after one-hot encoding for categorical (cat) variables. *Events* indicates the proportion of the non-censored events, *i.e.,* the events of interest for which $y_i = 1$. Missing indicates the proportion of missing entries in the $N \times d$ covariate matrix, and $t_{\max}$ is the time range for both censored and non-censored events. For all datasets except seer, that uses months, events are measured in days. In the experiments we do not convert time to a common scale and model it *as is*.

Details of the public datasets: flchain, support and seer, including preprocessing procedures, are provided in the above references. The other two datasets, ehr and sleep are not public but can be obtained upon request, see [7] and [53], respectively. For SLEEP we focus on the baseline clinical visit and aggregated demographics, medications and questionnaire data as covariates.

As shown in Table 1, survival datasets often contain substantial missingness, *e.g.* up to 23% in SEER data. Interestingly, [36] showed via the information-theoretic data processing inequality that there is no additional information to be gained by actively imputing missing values during training with an autoencoding arm, when compared to a simpler *pre-imputation* approach in which missing values are imputed with median and mode for continuous and categorical covariates, respectively. In view of this, here we adopt a pre-imputation strategy.

Quantitative evaluation

For a comprehensive quantitative evaluation of time-to-event models we consider three metrics that highlight different aspects of model performance: *i*) Concordance Index (C-Index) [22] to quantify preservation of pairwise orderings wrt ground truth events, *ii*) Coefficient of Variation (CoV) to assess uncertainty concentration by quantifying the dispersion of estimated time-to-event distributions, and *iii*) Calibration to asses the statistical consistency of the conditional survival distribution learned by a model relative to that of the ground truth. As previously discussed, a high-performing model is one that not only preserves pairwise ordering of event times, but also results in concentrated and well-calibrated time-to-event distributions. As discussed below, SFM outperforms other approaches in terms of calibration while being competitive in terms of C-Index (time ordering) and CoV (concentration).

Calibration—We evaluate calibration both visually and quantitatively. For the visual assessment, we plot the conditional survival distributions estimated from the model predictions using DKM in (3) and compare it with the empirical survival distribution (ground truth) using KM in (1), as shown in Figure 1. Alternatively, we plot the estimated conditional cumulative density function for each model using $1 - \hat{S}_{DKM}(t_i)$ against the marginal cumulative density function for the ground truth using $1 - \hat{S}_{KM}(t_i)$. In both cases,

$t_i \in \mathcal{T}$. If the estimated cumulative density matches the ground truth, the plotted curve will describe a diagonal line with unit slope. Curves above and below the diagonal *underestimate* and *overestimate* risk, respectively. Thus, for the quantitative assessment we calculate the *calibration slope,* which is obtained from the curve described by $1 - \hat{S}_{DKM}(t_i)$ *vs.* $1 - \hat{S}_{KM}(t_i)$. Since the cumulative density $F(t)$ is unknown for sampling-based approaches, e.g., DATE and SFM, we use a Gaussian Kernel Density Estimator (KDE) [46] on samples from the model, $\{t_{ns}\}_{s=1}^{200}$.

Results in Table 2 show that in terms of calibration slope fully nonparametric models, namely SFM and DATE, are better calibrated than S-CRPS and DRAFT, both parametrized as log-normally distributed models. Our approach is the best performing model across all datasets, followed by DATE, S-CRPS, CPH then DRAFT. We attribute these results to the fact that we directly match the survival function as part of model training. However, it is surprising that DATE and S-CRPS do not perform nearly as well considering that DATE adversarially matches the time-to-event distribution, thus indirectly matching the cumulative distribution, and S-CRPS that optimizes a proper scoring rule (the integral of Brier score at all possible thresholds [17]) that in principle should produce calibrated predictions.

For the ehr data it is not surprising that none of the models are well calibrated because observations in this dataset are not i.i.d. due to patients having multiple encounters. Since the models and KM-based estimators considered implicitly assume datasets are composed of i.i.d. observations, calibration does not necessarily hold. This necessitates further investigation, which we leave as interesting future work. However, to test the hypothesis that the model should be better calibrated in the i.i.d. case, we restricted the ehr dataset to the first encounter per patient (N=19,064), which results in a better calibrated SFM model (see the Supplementary Material).

Concordance Index—C-Index is arguably the most commonly used performance metric in survival analysis. This metric is useful to assess relative risk because it quantifies ordering rather than temporal accuracy. Models with high C-Index are good for the purpose of ranking observations into different risk categories, especially in a medical settings. Since the C-Index is evaluated on point estimates, we summarize time-to-event distributions as medians, *i.e.*, $\hat{t} = \text{median}(\{t_{ns}\}_{s=1}^{200})$, where t_{ns} is a sample from the trained model, $t_{ns} \sim G_{\boldsymbol{\theta}}(\boldsymbol{x}_n, \epsilon_s)$, on the test set.

Results in Table 2 show that none of the models has a clear advantage over the others, as the C-Index is largely comparable for the remaining four datasets. Apart from the small and high event rate support dataset where S-CRPS and DRAFT (both parametric log-normally distributed models) achieve (statistically) significantly higher C-Index compared to SFM (and CPH).

Coefficient of Variation—The Coefficient of Variation (CoV) quantifies the dispersion of a probability distribution. It is formally defined as $\sigma\mu^{-1}$, where σ and μ are respectively the standard deviation and mean of the distribution being tested. To summarize the variation of the time-to-event distributions estimated by different models on the test set, we use Mean

CoV, which is defined across all time-to-event predictions, *i.e,* $N_{\text{te}}^{-1}\sum_{n=1}^{N_{\text{te}}}\sigma_n\mu_n^{-1}$, where N_{te} is the size of the test set and σ_i and μ_i are sample standard deviations and means over $\{t_{ns}\}_{s=1}^{200}$. A model with concentrated time-to-event distributions is one for which mean CoV is as small as possible.

Figure 2 shows test set CoV distributions. We see that *i*) DRAFT and S-CRPS have considerably wider variation in CoV thus better 95% posterior coverage (see legend) compared to SFM and DATE; and *ii*) SFM and DATE are comparable, though DATE is slightly better. Note that we cannot evaluate CoV or coverage for CPH since in its standard form it only produces point estimates.

Table 2 shows that across all datasets DATE, SFM and S-CRPS are on average low-variance models while DRAFT is a considerably higher-variance model. DATE and SFM are the best-performing in terms predicting concentrated event times given that mean CoV < 0.5. High-variance time-to-event distributions are not desirable because when prediction uncertainty is large relative to the time range, they cannot be used to inform decision making. Examples of time-to-event distributions for all models as shown in the Supplementary Material.

Qualitative evaluation of calibration

There are several metrics for measuring the quality of calibration, *e.g.,* calibration slope and Brier score [37]. However, none of these summaries of calibration are as richly informative as visually comparing survival functions or cumulative density functions as described above. In Figure 2 we show calibration curves for two different datasets, seer and sleep, the largest and smallest dataset, respectively. See the Supplementary Material for figures corresponding to all other datasets including the conditional survival functions as in Figure 1. From these results (consistent across all datasets) we see that *i*) SFM performs better than the other approaches considered; *ii*) DRAFT is the worst performer; and *iii*) all approaches are poorly calibrated on seer data once half of the population has had events.

Under further examination of the seer data, we found there is a large subset of the population that gets administratively censored at $t = 80$ months (see the Supplementary Material), which explains the generalized sudden divergence of calibration in Figure 2. This type of *informative* censoring is not random and needs to be modeled appropriately. However, this extension is beyond the current scope and thus left as future work. Nonetheless, to test this idea, we truncated the data beyond $t = 88$ months and verified that the model is considerably better calibrated (see the Supplementary Material).

6 Conclusions

We have introduced a distribution-based Kaplan-Meier (DKM) estimator for evaluating calibration in time-to-event predictions. Leveraging this estimator, we introduced sFM, a survival-function- matched neural-network-based model for synthesizing calibrated time-to-event predictions. Our learning strategy matches the desired survival distribution without the need of an adversarial objective. Further, we showed that our survival distribution approach is related to earth mover's minimization. The proposed model outperforms other methods in estimating concentrated and calibrated time-to-event distributions, while

remaining competitive in terms of concordance index. As future work, we plan to extend the proposed approach to calibration in the non-i.i.d. setting, and to account for informative missingness.

A Coefficient of Variation Results

See figures 3, 4 and 5, for Coefficient of Variation (CoV) results.

B Calibration and Survival Function Results

The model calibration and survival plots for datasets SUPPORT, FLCHAIN, SLEEP, all EHR (non iid), subset EHR (iid) and SEER are shown in Figures 6, 7, 8, 9, 10 and 11 respectively.

B.1 SEER: Informative Censoring—

Figure 12 shows number of censoring and non-censored events over time. See Figure 13, for estimated calibration and survival function results for subset seer dataset when truncated at 88 months.

C Time-to-Event Distributions

Figures 14, 15, 16, 17 and 18, show the time-to-Event distributions heatmap over the time range $t_{\max}$.

D Batch Size Sensitivity Analysis

Table 3, shows SFM performance metrics across a range of batch sizes.

E Experimental Setup

In all experiments, SFM, DATE, DRAFT and S-CRPS are specified in terms of two-layer MLPs of 50 hidden units with Rectified Linear Unit (ReLU) activation functions, batch normalization [25] and apply dropout of $p = 0.2$ on all layers. We set the minibatch size to $M = 350$ and use the Adam [28] optimizer with the following hyperparameters: learning rate 3×10^{-4}, first moment 0.9, second moment 0.99, and epsilon 1×10^{-8}. We initialize all the network weights according to *Xavier* [16]. SFM and DATE inject noise in all layers, see [7] for more details. Datasets are split into training, validation and test sets as 80%, 10% and 10% partitions, respectively, stratified by non-censored event proportion. The validation set is used for early stopping and learning model hyperparameters. All models are trained using one NVIDIA P100 GPU with 16GB memory.

References

[1]. Alaa AM and van der Schaar M. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In NeurIPS, 2017.

[2]. Avati A, Duan T, Jung K, Shah NH, and Ng A. Countdown regression: Sharp and calibrated survival predictions. arXiv, 2018.

[3]. Bellot A and van der Schaar M. Boosted trees for risk prognosis. In MLHC, 2018.

[4]. Bender R, Augustin T, and Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, 2005.

[5]. Breslow N and Crowley J. A large sample study of the life table and product limit estimates under random censorship. The Annals of Statistics, 1974.

[6]. Brier GW. Verification of forecasts expressed in terms of probability. Monthey Weather Review, 1950.

[7]. Chapfuwa P, Tao C, Li C, Page C, Goldstein B, Carin L, and Henao R. Adversarial time-to-event modeling. ICML, 2018.

[8]. Chen L, Zhang Y, Zhang R, Tao C, Gan Z, Zhang H, Li B, Shen D, Chen C, and Carin L. Improving sequence-to-sequence learning via optimal transport. ICLR, 2019.

[9]. Cox DR. Regression models and life-tables. In Breakthroughs in statistics. 1992.

[10]. Dawid AP. The well-calibrated bayesian. Journal of the American Statistical Association, 1982.

[11]. DeGroot MH and Fienberg SE. The comparison and evaluation of forecasters. The statistician, 1983.

[12]. Dispenzieri A, Katzmann JA, Kyle RA, Larson DR, Therneau TM, Colby CL, Clark RJ, Mead GP, Kumar S, Melton LJ, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In Mayo Clinic Proceedings, 2012.

[13]. Fernández T, Rivera N, and Teh YW. Gaussian processes for survival analysis. In NeurIPS, 2016.

[14]. Fischl MA, Richman DD, Grieco MH, Gottlieb MS, Volberding PA, Laskin OL, Leedom JM, Groopman JE, Mildvan D, Schooley RT, et al. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. New England Journal of Medicine, 1987.

[15]. Fotso S. Deep neural networks for survival analysis based on a multi-task framework. arXiv, 2018.

[16]. Glorot X and Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.

[17]. Gneiting T and Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 2007.

[18]. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y. Generative adversarial nets. In NeurIPS, 2014.

[19]. Grathwohl W, Choi D, Wu Y, Roeder G, and Duvenaud D. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. ICLR, 2018.

[20]. Greenwood M et al. A report on the natural duration of cancer. A Report on the Natural Duration of Cancer., 1926.

[21]. Guo C, Pleiss G, Sun Y, and Weinberger KQ. On calibration of modern neural networks. ICML, 2017.

[22]. Harrell FE, Lee KL, Califf RM, Pryor DB, and Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics in medicine, 1984.

[23]. Hippisley-Cox J and Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. BMJ open, 2013.

[24]. Hosmer DW, Lemeshow S, and May S. Applied survival analysis. Wiley Blackwell, 2011.

[25]. Ioffe S and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[26]. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. JASA, 1958.

[27]. Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, and Kluger Y. Deep survival: A deep cox proportional hazards network. arXiv, 2016.

[28]. Kinga D and Adam JB. A method for stochastic optimization. In ICLR, 2015.

[29]. Kingma DP and Welling M. Auto-encoding variational bayes. ICLR, 2014.

[30]. Kleinbaum DG and Klein M. Survival analysis. Springer, 2010.

[31]. Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbiens N, et al. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. Annals of internal medicine, 1995.

[32]. Kolouri S, Park SR, Thorpe M, Slepcev D, and Rohde GK. Optimal mass transport: Signal processing and machine-learning applications. IEEE Signal Processing Magazine, 2017.

[33]. Kuleshov V, Fenner N, and Ermon S. Accurate uncertainties for deep learning using calibrated regression. ICML, 2018.

[34]. Lee C, Zame W, Alaa A, and Schaar M. Temporal quilting for survival analysis. In AISTATS, 2019.

[35]. Lee C, Zame WR, Yoon J, and van der Schaar M. Deephit: A deep learning approach to survival analysis with competing risks. AAAI, 2018.

[36]. Miscouridou X, Perotte A, Elhadad N, and Ranganath R. Deep survival analysis: Nonparametrics and missingness. In Machine Learning for Healthcare Conference, 2018.

[37]. Murphy AH. A new vector partition of the probability score. Journal of applied Meteorology, 1973.

[38]. Peterson AV Jr. Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. Journal of the American Statistical Association, 1977.

[39]. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, et al. The sleep heart health study: design, rationale, and methods. Sleep, 1997.

[40]. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 2018.

[41]. Ranganath R, Perotte A, Elhadad N, and Blei D. Deep survival analysis. In Machine Learning for Healthcare Conference, 2016.

[42]. Rezende DJ, Mohamed S, and Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. ICML, 2014.

[43]. Ries LAG, Young JL Jr, Keel GE, Eisner MP, Lin YD, and Horner M-JD. Cancer survival among adults: US SEER program, 1988–2001. Patient and tumor characteristics SEER Survival Monograph Publication, 2007.

[44]. Rubner Y, Tomasi C, and Guibas LJ. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000.

[45]. Salimans T, Zhang H, Radford A, and Metaxas D. Improving gans using optimal transport. ICLR, 2018.

[46]. Silverman BW. Density estimation for statistics and data analysis. Routledge, 2018.

[47]. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, and Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 2010.

[48]. Villani C. Optimal transport: old and new. Springer Science & Business Media, 2008.

[49]. Vinzamuri B, Li Y, and Reddy CK. Pre-processing censored survival data using inverse covariance matrix based calibration. IEEE Transactions on Knowledge and Data Engineering, 2017.

[50]. Wei L-J. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 1992.

[51]. Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 1992.

[52]. Yu C-N, Greiner R, Lin H-C, and Baracos V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In NeurIPS, 2011.

[53]. Zhang G-Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, and Redline S. The national sleep research resource: towards a sleep data commons. Journal of the American Medical Informatics Association, 2018.

[54]. Zhang Q and Zhou M. Nonparametric Bayesian lomax delegate racing for survival analysis with competing risks. In NeurIPS, 2018.

[55]. Zheng P, Yuan S, and Wu X. Safe: A neural survival analysis model for fraud early detection. AAAI, 2019.

[56]. Zhu X, Yao J, and Huang J. Deep convolutional neural network for survival analysis with pathological images. In Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, 2016.

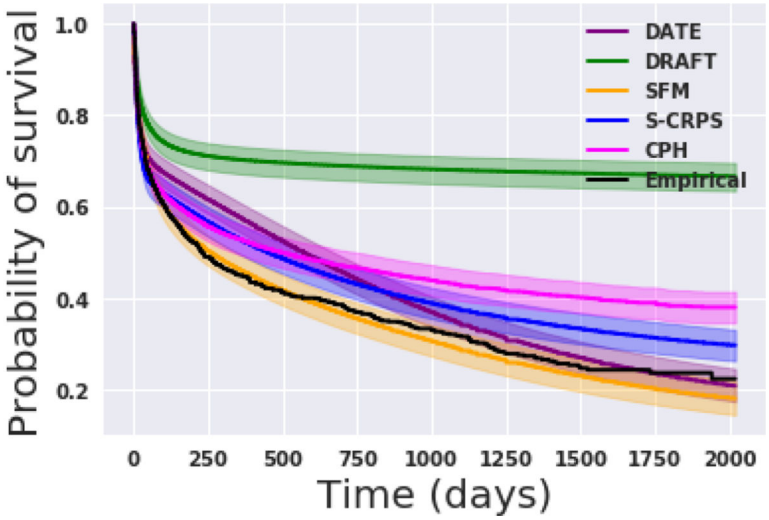

Figure 1:
Survival function estimates for SUPPORT data. Ground truth (Empirical) is compared to test set predictions from five models.

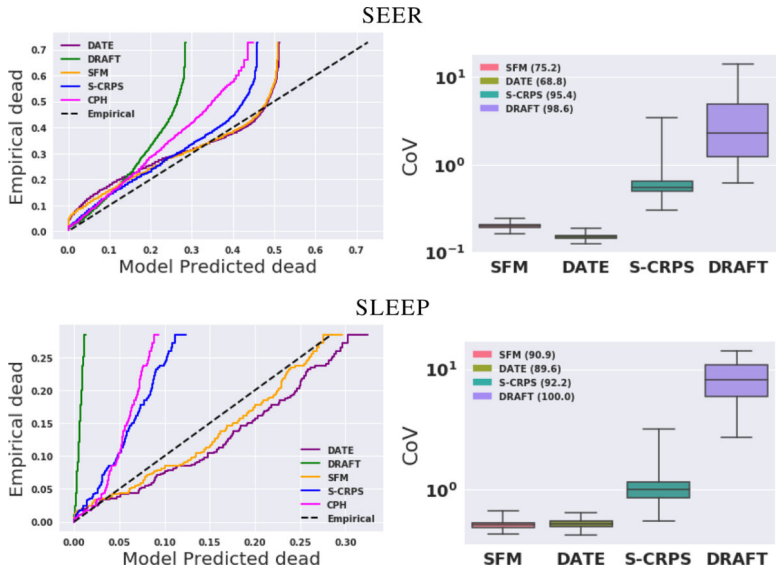

Figure 2:

Test set calibration and variation visualized for two datasets: seer and sleep (rows). Left: proportion of events of interest vs. predicted events. A perfectly calibrated model will follow the (dashed) diagonal line. Right: coefficient of variation (CoV) distributions. The legend shows the percentage of test set events covered by 95% intervals from predicted time-to-event distributions.

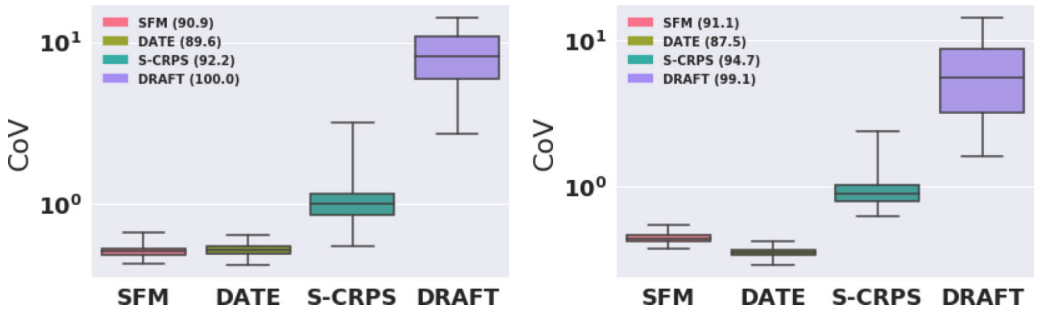

Figure 3:

Coefficient of Variation(CoV) distributions for (left) sleep and (right) flchain datasets. The legend shows the percentage of test set events covered by 95% intervals from predicted time-to-event distributions.

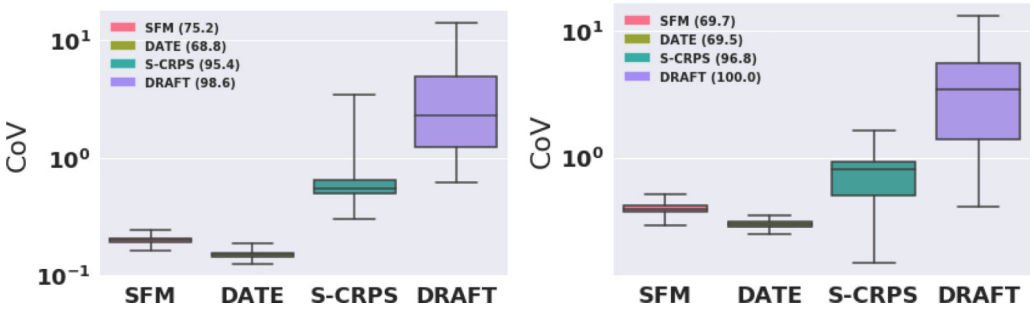

Figure 4:

Coefficient of Variation(CoV) distributions for (left) seer and (right) support datasets. The legend shows the percentage of test set events covered by 95% intervals from predicted time-to-event distributions.

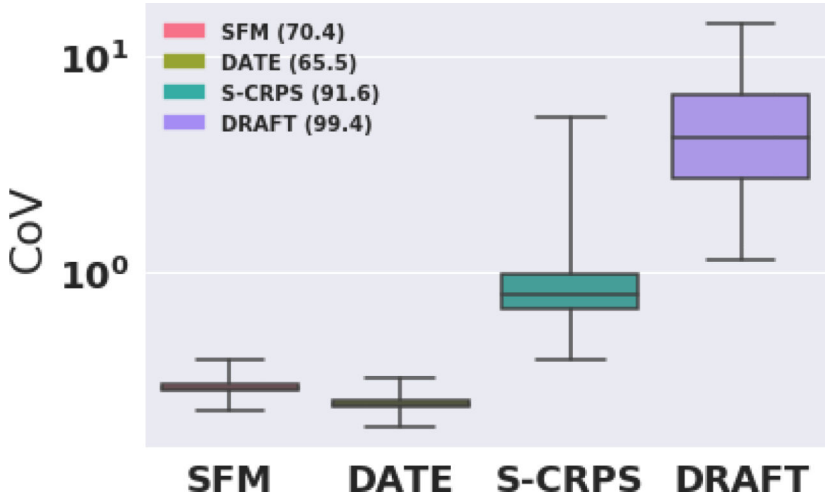

Figure 5:
Coefficient of Variation (CoV) distributions for ehr dataset. The legend shows the percentage of test set events covered by 95% intervals from predicted time-to-event distributions.

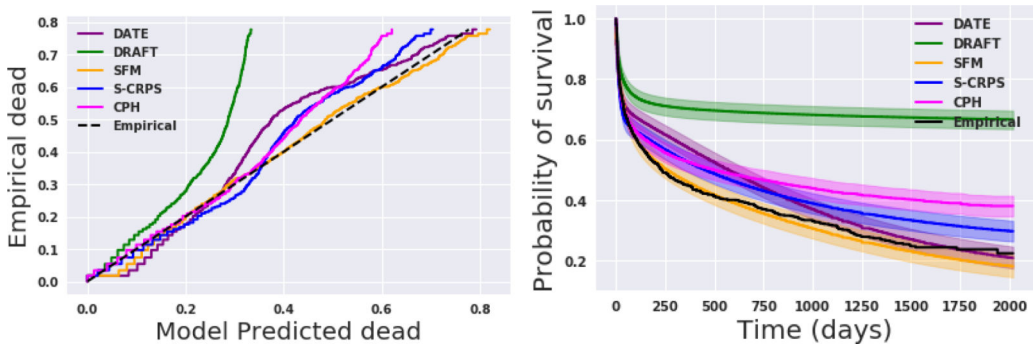

Figure 6:

Calibration (left) and Survival function estimates (right) for support data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model), S-CRPS and CPH).

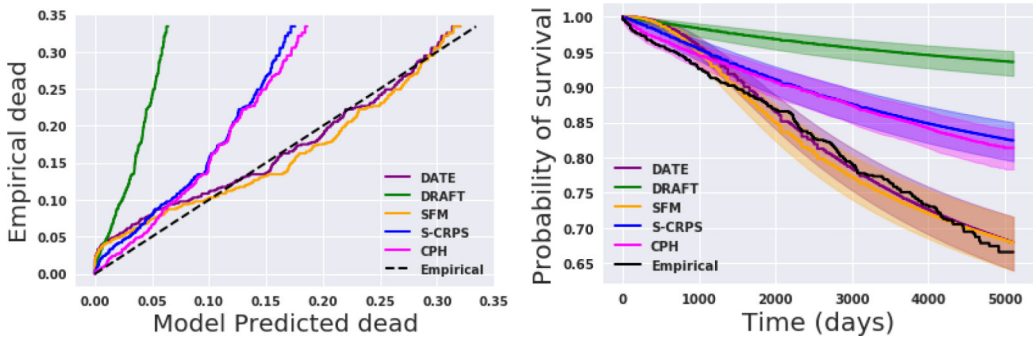

Figure 7:

Calibration(left) and Survival function estimates (right) for flchain data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (proposed model), S-CRPS and CPH).

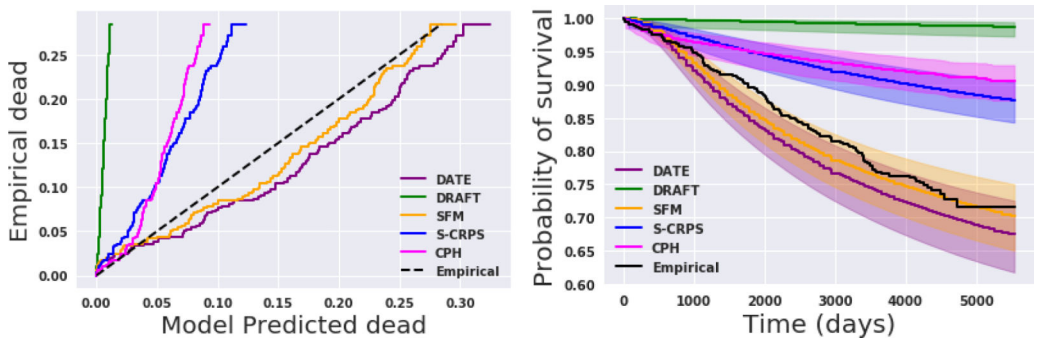

Figure 8:

Calibration (left) and Survival function estimates (right) for sleep data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model) S-CRPS, and CPH).

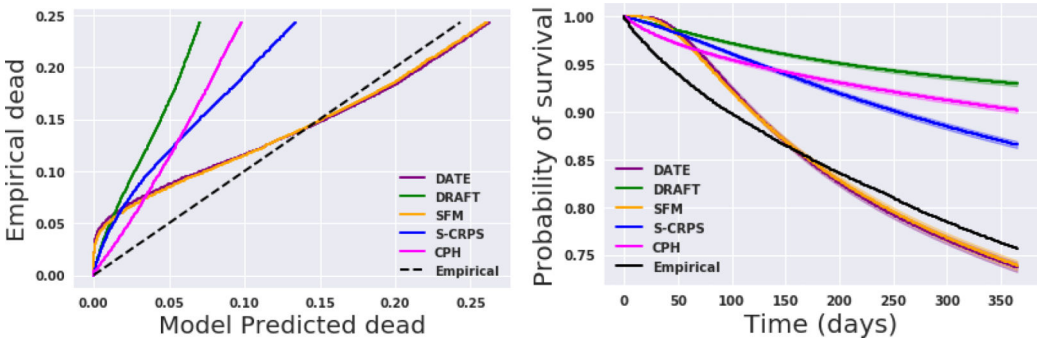

Figure 9:

Calibration (left) and Survival function estimates (right) for ehr all (non iid) data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model), S-CRPS, and CPH).

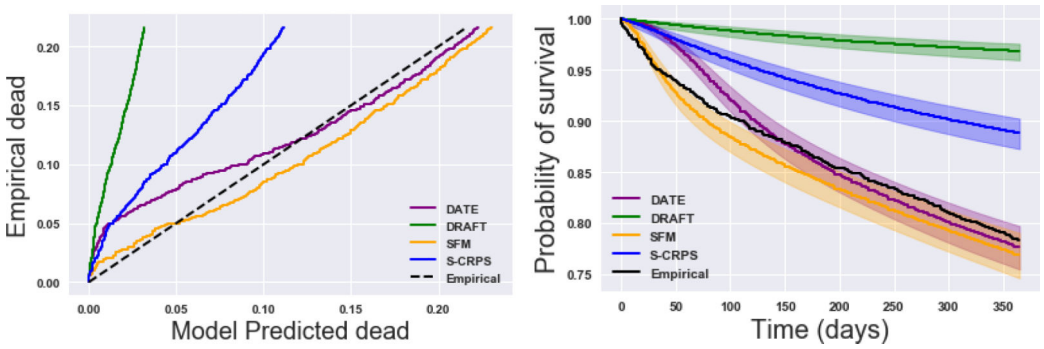

Figure 10:

Calibration (left) and Survival function estimates (right) for ehr subset (iid) data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model) and S-CRPS).

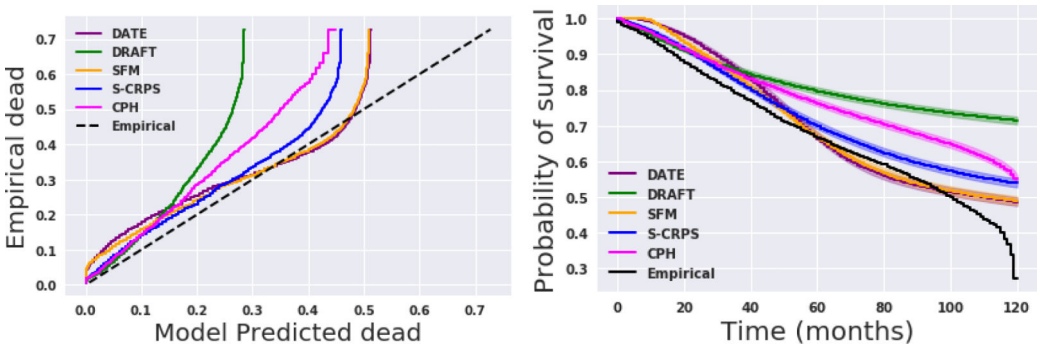

Figure 11:

Calibration(left) and Survival function estimates (right) for seer data. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model), S-CRPS and CPH).

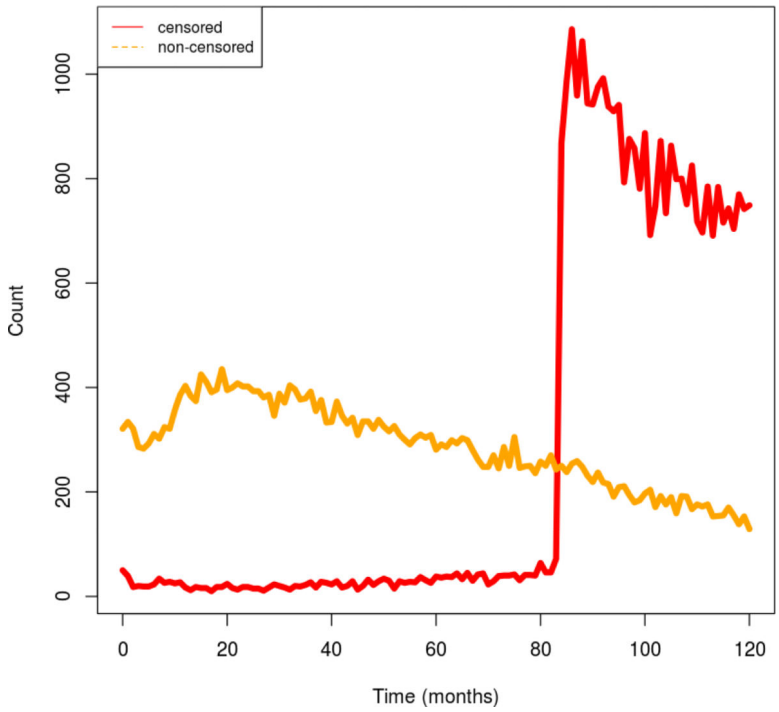

Figure 12:
Count of censored and non-censored events as a function of time for SEER data.

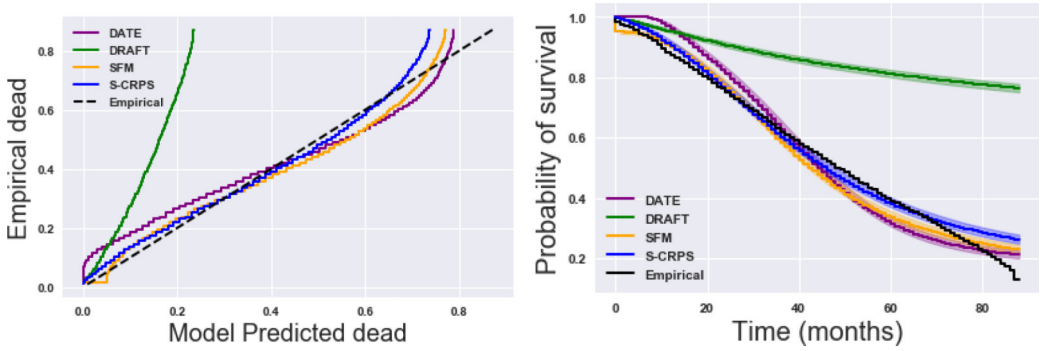

Figure 13:

Calibration(left) and Survival function estimates (right) for subset seer data truncated at 88 months. Ground truth (Empirical) is compared to predictions from four models (DATE, DRAFT, SFM (our proposed model) and S-CRPS).

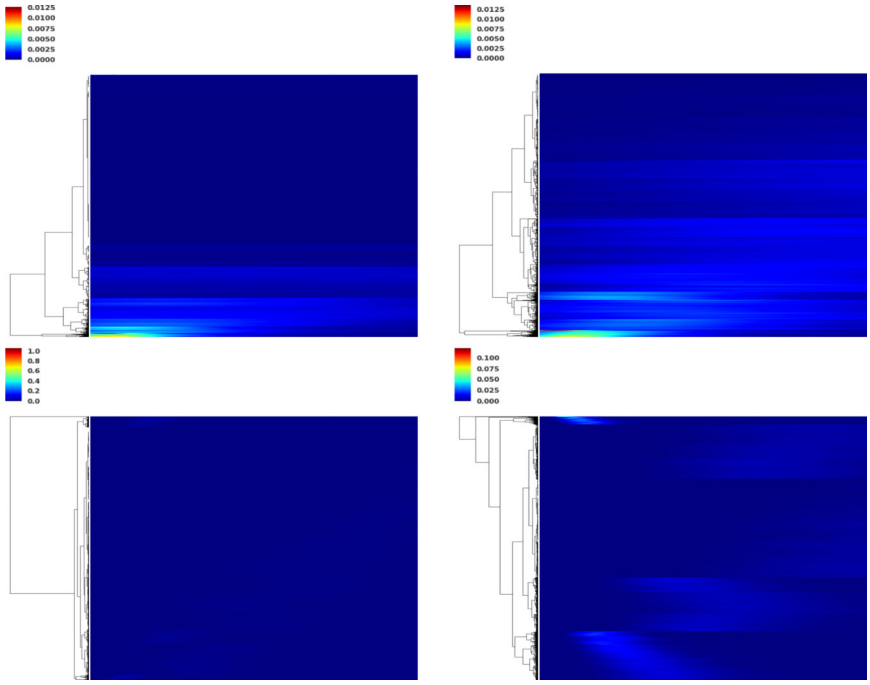

Figure 14:
Heatmap of time-to-event distributions on EHR data for DRAFT (top-left), S-CRPS (top-right), DATE (bottom-left) and SFM (bottom-right). The x-axis is the time range $t_{\max}$.

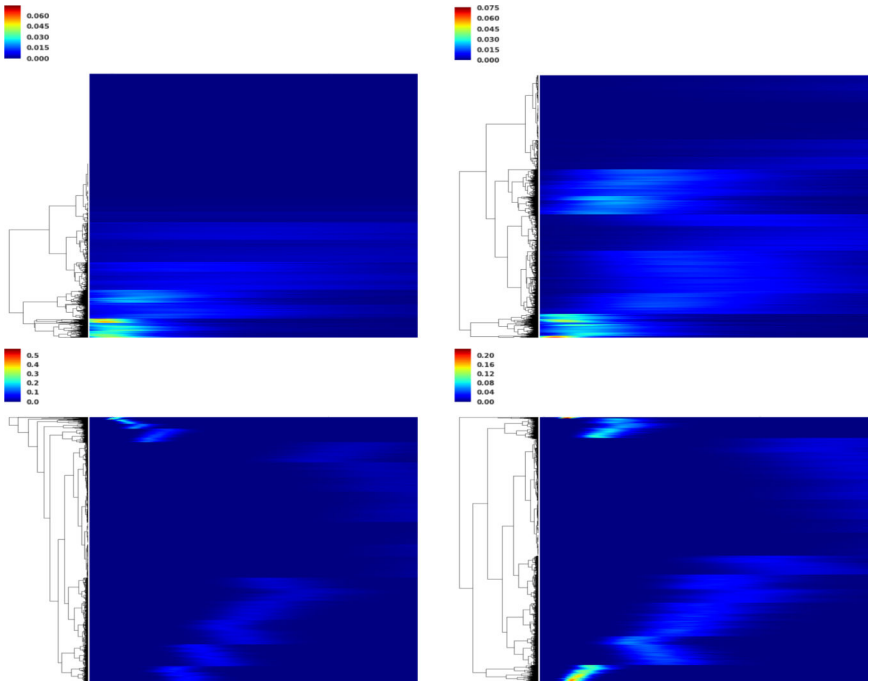

Figure 15:
Heatmap of time-to-event distributions on SEER dataset for DRAFT (top-left), S-CRPS (top-right), DATE (bottom-left) and SFM (bottom-right). The x-axis is the time range $t_{\max}$.

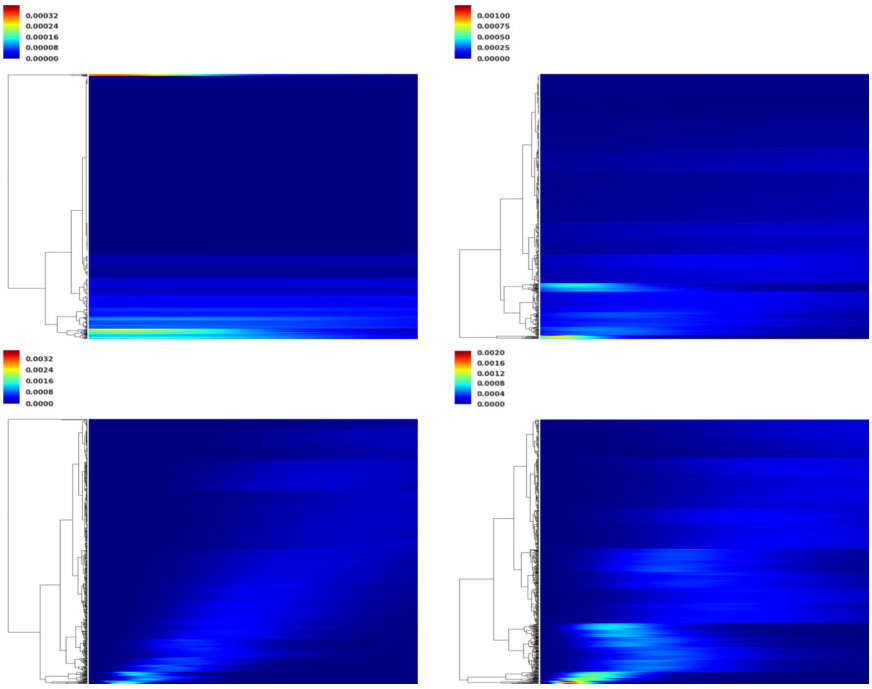

Figure 16:

Heatmap of time-to-event distributions on FLCHAIN data for DRAFT (top-left), S-CRPS (top-right), DATE (bottom-left) and SFM (bottom-right). The x-axis is the time range $t_{\max}$.

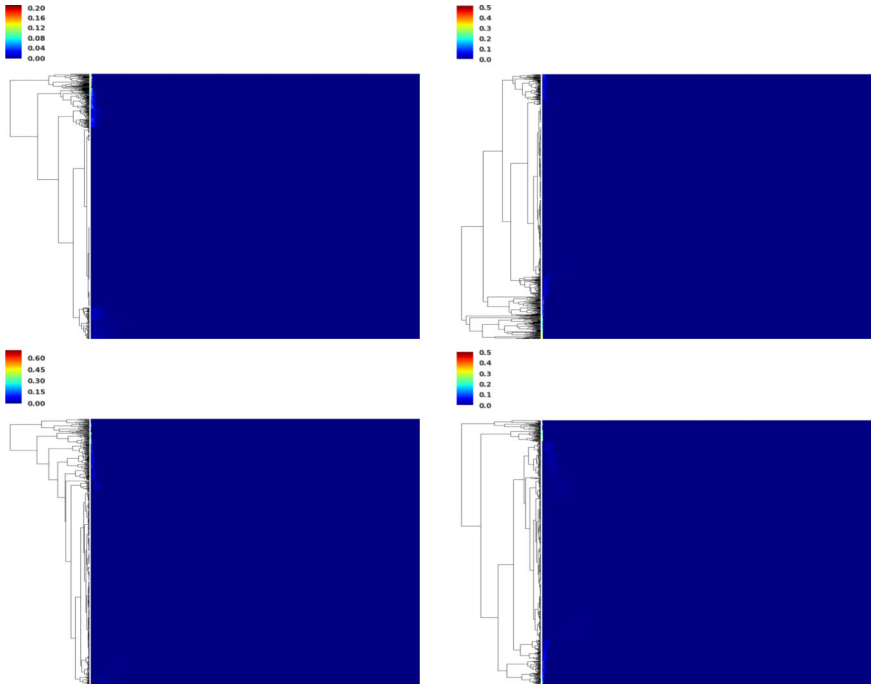

Figure 17:

Heatmap of time-to-event distributions on SUPPORT data for DRAFT (top-left), S-CRPS (top-right), DATE (bottom-left) and SFM (bottom-right). The x-axis is the time range $t_{\max}$.

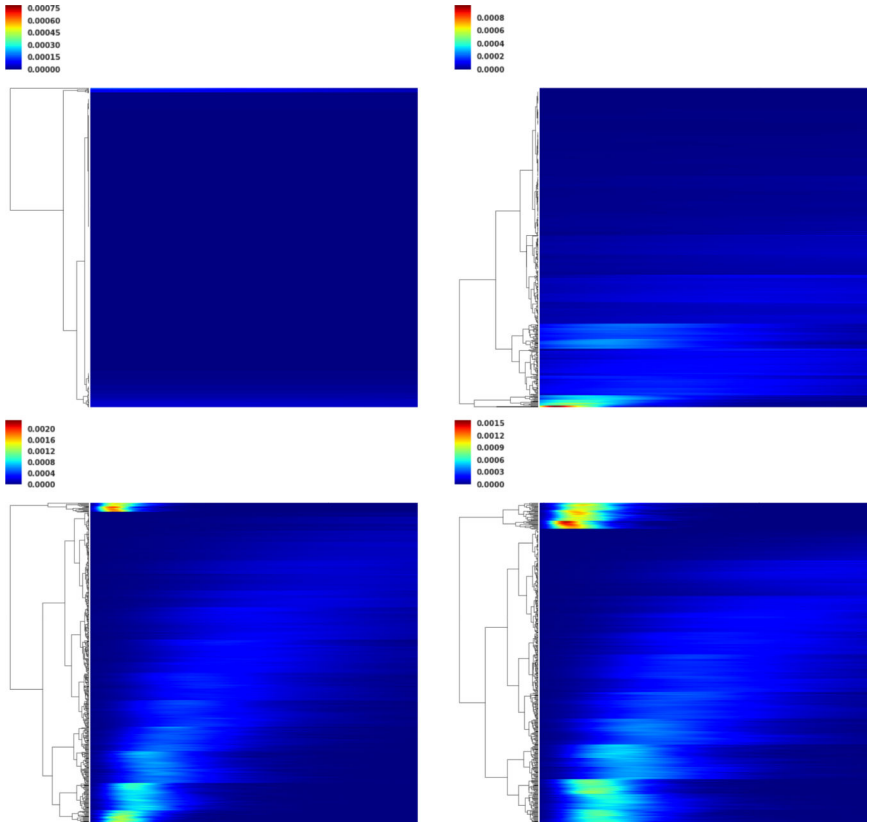

Figure 18:

Heatmap of time-to-event distributions on SLEEP data for DRAFT (top-left), S-CRPS (top-right), DATE (bottom-left) and SFM (bottom-right). The x-axis is the time range $t_{\max}$.

Table 1:

Summary statistics of the datasets used in the experiments. The time range, $t_{\max}$, is noted in days except for seer for which time is measured in months.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP
Events (%)	23.9	27.5	68.1	51.0	23.8
N	394,823	7,894	9,105	68,082	5026
d (cat)	729 (106)	26 (21)	59 (31)	789 (771)	206
Missing (%)	1.9	2.1	12.6	23.4	18.2
$t_{\max}$	365	5,215	2,029	120	5,794

Table 2:

Performance metrics. SFM is the proposed model.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP
Calibration slope					
DATE	0.7537	0.9668	0.9068	0.9161	0.9454
DRAFT	3.2138	5.4183	2.9640	2.0763	25.2855
S-CRPS	1.6246	1.9662	1.1795	1.1613	2.5746
CPH	2.5543	1.9116	1.3909	1.4358	3.8278
SFM	**0.7734**	**0.9807**	**0.9405**	**0.9540**	**1.0235**
Mean CoV					
DATE	**0.2477**	**0.3585**	**0.2987**	**0.1485**	0.5168
DRAFT	5.0305	6.2952	3.8689	3.4501	8.4918
S-CRPS	0.8585	0.9412	0.7351	0.6036	1.0240
CPH	-	-	-	-	-
SFM	0.2953	0.4484	0.3930	0.1993	**0.5045**
C-Index					
DATE	0.7756	0.8264	0.8421	**0.8320**	0.7416
DRAFT	**0.7796**	0.8341	0.8560	0.8310	**0.7617**
S-CRPS	0.7704	0.8286	**0.8685**	0.8298	0.7529
CPH	0.7542	**0.8344**	0.8389	0.8223	0.6435
SFM	0.7786	0.8318	0.8319	0.8314	0.7491

Table 3:

SFM batch size sensitivity on flchain dataset.

	100	**250**	**500**	**750**	**1000**
Calibration slope					
	1.0110	0.9766	0.9807	0.9864	**0.9916**
Mean CoV					
	0.4740	**0.4026**	0.4484	0.4332	0.4672
C-Index					
	0.8302	0.8294	**0.8318**	0.8296	0.8287