Survival Cluster Analysis

Paidamoyo Chapfuwa
Duke University
paidamoyo.chapfuwa@duke.edu

Chunyuan Li
Microsoft Research, Redmond
chunyl@microsoft.com

Nikhil Mehta
Duke University
nm208@duke.edu

Lawrence Carin
Duke University
lcarin@duke.edu

Ricardo Henao
Duke University
ricardo.henao@duke.edu

ABSTRACT

Conventional survival analysis approaches estimate risk scores or individualized time-to-event distributions conditioned on covariates. In practice, there is often great population-level phenotypic heterogeneity, resulting from (unknown) subpopulations with diverse risk profiles or survival distributions. As a result, there is an unmet need in survival analysis for identifying subpopulations with distinct risk profiles, while jointly accounting for accurate individualized time-to-event predictions. An approach that addresses this need is likely to improve characterization of individual outcomes by leveraging regularities in subpopulations, thus accounting for population-level heterogeneity. In this paper, we propose a Bayesian nonparametrics approach that represents observations (subjects) in a clustered latent space, and encourages accurate time-to-event predictions and clusters (subpopulations) with distinct risk profiles. Experiments on real-world datasets show consistent improvements in predictive performance and interpretability relative to existing state-of-the-art survival analysis models.

CCS CONCEPTS

• **Computing methodologies → Machine learning algorithms**; **Machine learning**; **Machine learning approaches**; *Learning latent representations*;

KEYWORDS

survival analysis, Bayesian nonparametrics, clustering, time-to-event

1 INTRODUCTION

Time-to-event models have primarily focused on either estimating a (point estimate) risk score or individualized time-to-event distributions. Parametric models estimate the time-to-event distribution

conditional on covariates by assuming a parametric form of the event distribution, *i.e*, exponential, Weibull, log-normal, *etc*. Parametric models fall under the Accelerated Failure Time (AFT) [48] framework, provided they assume covariates either accelerate or decelerate the time-to-event. Assuming a parametric distribution is inflexible as the hazard function depends on the selected baseline distribution, for example assuming an exponential distribution, yields a constant hazard rate function. Alternatively, Cox Proportional Hazards (CoxPH) [14], a semi-parametric, linear model for estimating relative risks is widely used in practice, as it does not require one to specify the baseline distribution. For pre-specified time-horizons, the non-parametric Random Survival Forest (RSF) [26] was proposed to estimate the cumulative hazard function based on an ensemble of binary decision trees, albeit often limited by scaling problems for large and high-dimensional datasets.

With recent advances in machine learning, deep learning methods have improved classical survival analysis methods by leveraging non-linear relationship between covariates, for improved time-to-event or risk score predictions. Deep learning methods inspired by CoxPH or AFT have been proposed, *e.g.*, DeepSurv [29], Deep Survival Analysis (DSA) [42], Deep Regularized Accelerated Failure Time (DRAFT) [11], Gaussian-process-based models [4, 19], and the Survival Continuous Ranked Probability Score (S-CRPS) [5]. Sampling-based nonparametric methods have been proposed as well, *e.g.*, normalizing-flow-based DSA [37], adversarial-learning-based Deep Adversarial Time to Event (DATE) [11] and Survival Function Matching (SFM) [12]. Another class of nonparametric methods discretize time-to-event to predict survival probability within pre-specified discrete-interval event times with logistic-regression-based methods [20, 33, 51]. Further, deep learning methods have also successfully addressed calibration [5, 12, 34] and competing risks [4, 52].

Clustering based on risk-profiles in survival analysis is relatively under-explored in machine learning, but is critical in applications such as (clinical) decision making. Identifying phenotypically heterogeneous subpopulations in the context of risk prediction is an important step toward machine-learning-based models for precision medicine [13, 16]. Existing clustering methods for stratifying risks in survival analysis include feature based K-means (see Figure 1(a)) or hierarchical clustering [2, 17, 46]. Principal component cluster analysis has also been considered [3]. However, it is well understood that feature-based clustering in covariate space may produce clusters that are not consistent with survival outcomes [6, 21], particularly for high-dimensional datasets, such as gene expression data.
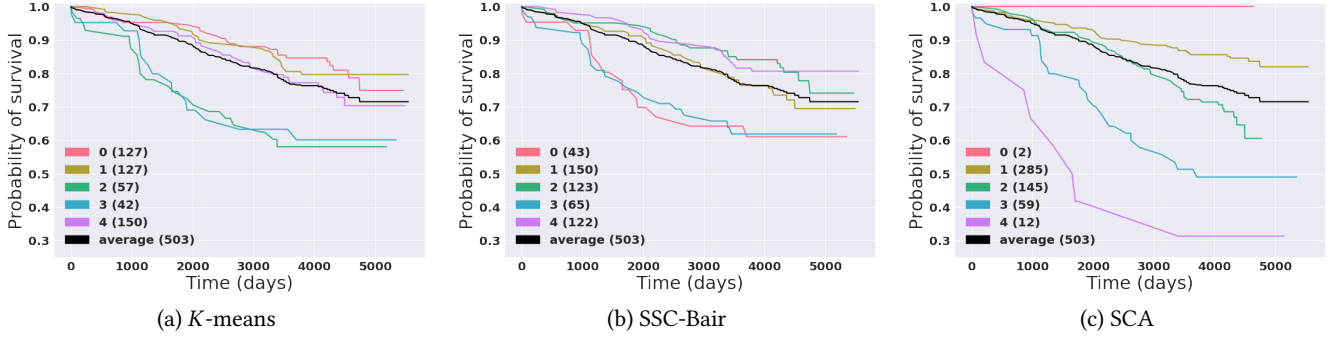
(a) K-means (b) SSC-Bair (c) SCA

Figure 1: Cluster-specific Kaplan-Meier survival profiles for three clustering methods on the SLEEP dataset (See Section 4 for details). (a) Standard K-means. (b) CoxPH-based covariate selection followed by K-means. (c) Proposed approach for joint learning of individualized time-to-event predictions and clustering. By jointly learning clustering with respect to both the covariates x and predicted time-to-event t, our model (SCA) can identify high-, medium- and low-risk individuals. Demonstrating the need to account for time information via a non-linear transformation of covariates when clustering survival datasets.

Methods that account for survival outcomes in clustering include CoxPH-inspired techniques [6, 21], implemented as a two-step process: first, high CoxPH scoring covariates are selected, then a classical clustering approach like K-means is applied (see Figure 1(b)). However, CoxPH-based approaches are limited by the proportional hazards assumption. Alternatively, Xia et al. [49] proposed an outcome driven attention-based multi-task deep learning model for classification and then applied K-means on the latent representations to cluster subjects with acute coronary syndrome. More recently, Mouli et al. [38] introduced DeepCLife, a method that learns clusters by maximizing the pairwise differences between the survival functions of all cluster pairs. This is done by indirectly maximizing the logrank score [36]. Unlike DeepCLife, which aims to optimize clusters but not predictions, our goal is to jointly characterize time-to-event predictive distributions from a clustered latent space conditioned on covariates (see Figure 1(c)).

We propose a model for time-to-event predictions equipped with a structured latent representation that allows for clustering via a prior for infinite mixture of distributions. We circumvent the challenges associated with infinite mixtures in stochastic learning by leveraging a truncated Dirichlet process (DP) with a stick breaking representation. The proposed model, termed Survival Clustering Analysis (SCA), is specified as: *i*) a deterministic encoder that maps covariates into a latent representation; *ii*) a stochastic time-to-event predictor that feeds from the latent representation; and *iii*) a distribution matching objective that encourages latent representations to behave as a mixture of distributions following a DP structure. This approach allows identification and analysis of phenotypically heterogeneous subpopulations. Our experiments demonstrate that SCA yields consistent improvements in predictive performance and cluster quality relative to existing methods.

2 BACKGROUND

In a conventional time-to-event (survival analysis) setup, we are given N observations. Individual observation are described by triplets $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, l_i)\}_{i=1}^N$, where $\boldsymbol{x}_i = \{x_i, ..., x_p\}$ is a p-dimensional vector of covariates, t_i is the time-to-event and $l_i \in \{0, 1\}$ is the censoring indicator. When $l_i = 0$ (censored) the subject has *not*
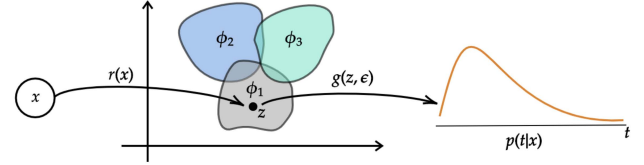

Figure 2: Illustration of Survival Clustering Analysis (SCA). The latent space has a mixture-of-distributions structure, illustrated as three mixture components $\{\phi_k\}_{k=1}^3$. Observation x is mapped into its latent representation via a deterministic encoding $z = r_\psi(x)$ belonging to ϕ_1, which is then used to stochastically predict (via sampling) the time-to-event via $t = g_\theta(z, \epsilon)$.

experienced an event up to time t_i, while $l_i = 1$ indicates observed (ground truth) event times.

Time-to-event models are conditional on covariates: the event time density function $f(t|\boldsymbol{x})$, the hazard rate (risk score) function $h(t|\boldsymbol{x})$ or the survival function $S(t|\boldsymbol{x}) = P(T > t) = 1 - F(t|\boldsymbol{x})$, also known as the probability of failure occurring after time t, where $F(t|\boldsymbol{x})$ is the cumulative density function. From standard survival function definitions [31], the relationship between these three characterizations is formulated as $f(t|\boldsymbol{x}) = h(t|\boldsymbol{x})S(t|\boldsymbol{x})$.

In practice, modern (often large) datasets are not homogeneous but composed of phenotipically heterogeneous subpopulations, *i.e.*, subsets of observations that cluster according to both covariates and time-to-event similarities. In a clinical setting for instance, identification of, *e.g.*, high-, medium- and low-risk subpopulations that are equipped with accurate estimates of time-to-event has the potential to result in a more cost effective way of targeting interventions, treatments or care delivery. We formulate an approach to jointly learn individualized time-to-event distributions and clusters informed by time-to-event profiles.

3 SURVIVAL CLUSTER ANALYSIS

The Bayesian nonparametrics approach formulated below encourages latent representations to behave as a mixture of distributions,

following a Dirichlet Process (DP) structure via a distribution matching approach. Further, we learn to cluster the latent space in a stochastic manner for which the number of clusters is unknown. To demonstrate the efficacy of our clustering algorithm, we also present a time-to-event prediction formulation, leveraging current state-of-the-art time-to-event prediction models. See the Supplementary Material for the list of variable definitions used in our formulation.

3.1 Clustering with Dirichlet Process

A DP is formally defined as $G \sim \text{DP}(\gamma_o, G_o)$ and parametrized by the *base probability measure* G_o and *concentration parameter* $\gamma_o > 0$ [18]. With probability one [45]:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \qquad \pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l), \tag{1}$$

where $\phi_k \sim G_o$, δ_{ϕ_k} represents a probability measure concentrated at ϕ_k and $V_k \sim \text{Beta}(1, \gamma_o)$ are *stick-breaking* weights with statistics that depend on parameter γ_0. The sequence $\boldsymbol{\pi} = \{\pi_k\}_{i=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$, with probability one, such that $\boldsymbol{\pi} \sim \text{GEM}(\gamma_o)$ [40]. Further, note that π_k represents the likelihood that $G = \phi_k$.

Discrete distribution G is suitable as a prior for mixture components in infinite mixture models [43]. Further, the stick-breaking process [45] that generates $\boldsymbol{\pi}$ results in a mechanism that allows one to learn the number of mixture components (clusters) from data. In fact, the number of distinct atoms, $\{\phi_k\}_{k=1}^{\infty}$, has been shown to grow with the size of the data as $O(\log N)$ [1]. So motivated, we assume that data embedded in a latent space are distributed according to a mixture of distributions with parameters specified by the *base probability measure* G_o, as described below.

Assuming exchangeable latent representations $\{z\}_{i=1}^{N}$, we propose generating event times following the generative process below

$$p(c) = \sum_{k=1}^{\infty} \pi_k \delta_{c_k} \tag{2}$$

$$z_n \sim st(c_{u_n}, \nu) \tag{3}$$

$$t_n \sim g_{\boldsymbol{\theta}}(z_n, \epsilon_n), \tag{4}$$

where $g_{\boldsymbol{\theta}}(z, \epsilon)$ is a function that implicitly represents the conditional time-to-event density, $f(t|x)$, specified as a neural network with parameters $\boldsymbol{\theta}$. The source of stochasticity, ϵ, for $g_{\boldsymbol{\theta}}(z, \epsilon)$, is set to a simple distribution $\epsilon \sim p_{\epsilon}$, *e.g*, uniform or Gaussian. The latent representation for the n-th observation, z_n is distributed according to $\phi_{u_n} = st(c_{u_n}, \nu)$, where u_n is the mixture component membership indicator for z_n. Lastly, together with (3), $p(c)$ in (2) represents an infinite mixture of Student's t-distributions with ν degrees of freedom and means $\{c_k\}_{k=1}^{\infty}$, each of which is drawn independently from the base probability measure G_o as $c_k \sim \mathcal{N}(0, I)$.

The Student's t distribution in (3) is a general yet parametrically simple distribution, robust to outliers and amenable to efficient computations and gradient estimates. It has been widely used in machine learning for mixture modeling [43], clustering [50] and visualization [35]. Further, we formulate the t distribution according to the normal-inverse-gamma likelihood, where marginalizing out the variance yields a Student-t distribution, see [8] for details.

Interestingly, as special cases, when $\nu = 1$, z_n is Cauchy distributed while for $\nu > 3$, z_n approaches a Gaussian distribution.

The generative process above further requires learning a mapping function from covariates to latent space, $z_n = r_{\boldsymbol{\psi}}(x_n)$ with parameters $\boldsymbol{\psi}$, that is globally consistent with the mixture model prior in (2) and (3), parameterized by $\{\pi_k, c_k\}_{k=1}^{\infty}$. In addition, we also need to learn the parameters $\boldsymbol{\theta}$ of the time-to-event generating function $g_{\boldsymbol{\theta}}(z_n, \epsilon_n)$ in (4). This specification, illustrated in Figure 2, constitutes the proposed Survival Clustering Analysis (SCA).

Note that unlike existing unsupervised and supervised autoencoding approaches [27, 30, 47], we do not seek to model the covariates, x. Rather, we make time-to-event predictions based on a latent representation specified as a function of observed covariates, required to be consistent with a mixture of distributions prior. Consequently, we need not specify a decoding arm to reconstruct the covariates, x.

In practice, learning the mixture component assignments u_n and a potentially infinite number of mixture components with minibatches (stochastically) is challenging, because the former constitutes a discrete random variable and the latter requires keeping track of the number of non-empty mixture components during learning. To circumvent this, we learn the mixture component assignments probabilistically as $q(u_n = k|x_n)$, and use a truncated representation of the DP formulation [9, 25], which for large enough truncation number, denoted as K, is virtually indistinguishable from a standard DP [25].

3.2 Latent-Space Representation

Following the conventional maximum likelihood formulation for mixture models [8], we can approximate the distributions for the mixture assignments and mixture proportions as follows

$$q(u_n = k|x_n) = \frac{\alpha_{nk}}{\sum_{k=1}^{K} \alpha_{nk}}$$

$$\alpha_{nk} \propto \pi_k p(r_{\boldsymbol{\psi}}(x_n)|c_k, \nu)$$

$$q(\boldsymbol{\pi}|\xi, \{x_1\}_{n=1}^{M}) = \text{Dir}(\boldsymbol{\pi}|\xi) \tag{5}$$

$$\xi_k = \frac{1}{K} + \sum_{n=1}^{M} q(u_n = k|x_n),$$

where $\boldsymbol{\xi} = \{\xi_k\}_{k=1}^{K}$ is a variational parameter for expectation $\mathbb{E}[\boldsymbol{\pi}]$, M is the minibatch size and we have replaced z_n in (3) with the encoding of covariates into latent space, *i.e.* $z_n = r_{\boldsymbol{\psi}}(x_n)$. However, (5) is not necessarily consistent with the DP in (2) and its stick-breaking prior, $\boldsymbol{\pi} \sim \text{GEM}(\gamma_0)$, which from (1) should result in

$$p(u_n = k|x_n) = \frac{\beta_{nk}}{\sum_{k=1}^{K} \beta_{nk}}$$

$$\beta_{nk} \propto V_k \prod_{l=1}^{k-1} (1 - V_l) p(r_{\boldsymbol{\psi}}(x_n)|c_k, \nu)$$

$$p(\boldsymbol{\pi}|\boldsymbol{\gamma}, \{x_n\}_{n=1}^{M}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\gamma}) \tag{6}$$

$$\gamma_k = \gamma_0 + \sum_{n=1}^{M} p(u_n = k|x_n),$$

where $V_k \sim \text{Beta}(1, \gamma_0)$, which in practice is complicated by the need to sample from the mixture proportion weights $\{V_k\}_{k=1}^{K}$. In

our implementation, instead of sampling from V_k, we use its expectation, i.e., $\mathbb{E}[V_k] = (1 + \gamma_0)^{-1}$. Alternatively, we could also use a reparameterizable distribution such as the Kumaraswamy distribution, which is closely related to the Beta distribution as in [39]. However, we found that using expectations, which is common in variational formulations [9, 28], works well in practice.

In order to make $q(\boldsymbol{\pi}|\xi, \{\boldsymbol{x}_1\}_{n=1}^{M})$ in (5) and $p(\boldsymbol{\pi}|\gamma, \{\boldsymbol{x}_n\}_{n=1}^{M})$ in (6) consistent, we want their distributions to match, i.e., we seek to learn $\boldsymbol{\psi}$ of $\boldsymbol{z}_n = r_\psi(\boldsymbol{x}_n)$, $\{\pi_k\}_{k=1}^{K}$ and $\{\boldsymbol{c}_k\}_{k=1}^{K}$, so the approximation $q(\boldsymbol{\pi}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ matches the desired stick breaking behavior of (6). For this purpose, we minimize

$$\ell_{\mathrm{dp}}(\boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^{K}; \mathcal{D}) = \\ \mathrm{KL}\left(q(\boldsymbol{\pi}|\xi, \{\boldsymbol{x}_n\}_{n=1}^{M})||p(\boldsymbol{\pi}|\gamma, \{\boldsymbol{x}_n\}_{n=1}^{M})\right). \quad (7)$$

The KL Divergence between the two Dirichlet distributions q and p with respect to their corresponding parameters ξ and γ, has a desirable closed form formulation defined as

$$\mathrm{KL}\left(q||p)\right) = \ln\Gamma(\xi_0) - \ln\Gamma(\gamma_0) + \sum_{k=1}^{K}\left(\ln\Gamma(\gamma_k) - \ln\Gamma(\xi_k)\right) \\ + \sum_{k=1}^{K}(\xi_k - \gamma_k)\left(\Phi(\xi_k) - \Phi(\xi_0)\right), \quad (8)$$

where $\xi_0 = \sum_{k=1}^{K}\xi_k$, $\gamma_0 = \sum_{k=1}^{K}\gamma_k$, $\Phi(\cdot)$ is the *digamma function* and $\Gamma(\cdot)$ is the *Gamma function*.

This loss function is used during learning to update $\boldsymbol{\psi}$ and $\{\boldsymbol{c}_k\}_{k=1}^{K}$. For $\{\pi_k\}_{k=1}^{K}$, the mixture proportions, we use a simple updating procedure akin to online expectation-maximization (EM) [10]. In particular, we update iteratively as

$$\boldsymbol{\pi}^{t+1} = \eta\boldsymbol{\pi}^t + (1 - \eta)\mathbb{E}[q(\boldsymbol{\pi}|\xi, \{\boldsymbol{x}_n\}_{n=1}^{M})], \quad (9)$$

where $0 < \eta < 1$ is the step size and we initialize $\boldsymbol{\pi}_k^0 = 1/K$. In practice, we set $\eta = 0.9$; however, η can also be selected using grid search. The online approach in (9) is widely used to update global parameters in stochastic learning procedures. For instance, it has been used to learn the population mean and variance in batch normalization [24].

3.3 Time-to-Event Distributions

In addition to the clustered, mixture representation of the latent space, we also seek a high-performing time-to-event model that yields concentrated, accurate and calibrated time-to-event predictions, while accounting for censored event times ($l_n = 0$). We borrow the accuracy objective from DATE [11] and the calibration objective from SFM [12]. Below we describe these objectives in the context of our formulation.

Accuracy Objective. The dataset $\mathcal{D}$ is split into two disjoint sets $(t, \boldsymbol{x}) \sim p_c$ and $(t, \boldsymbol{x}) \sim p_{nc}$, where p_c and p_{nc} represent censored and non-censored empirical distributions for these sets, respectively. We leverage the accuracy objective from DATE [11] formulated as

$$\ell_{\mathrm{acc}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) = \mathbb{E}_{(t,\boldsymbol{x}) \sim p_c, \boldsymbol{\epsilon} \sim p_\epsilon}[\max\left(0, t - g_\theta(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon})\right)] \\ + \mathbb{E}_{(t,\boldsymbol{x}) \sim p_{nc}, \boldsymbol{\epsilon} \sim p_\epsilon}[||t - g_\theta\left(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon}\right)||_1], \quad (10)$$

where $\boldsymbol{\epsilon} \sim p_\epsilon$ has a simple distribution (uniform or Gaussian). $\ell_{\mathrm{acc}}(\boldsymbol{\theta}; \mathcal{D})$ encourages that time-to-event samples from the model,

evaluated on censored observations, $l_n = 0$, are larger than the censoring time, while close to the ground truth for non-censored (observed) events, $l_n = 1$.

Calibration Objective. We desire that samples generated from the model $g_\theta(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon})$ match the empirical marginal distribution $p(t)$. We borrow the calibration objective from SFM [12] defined over the set of distinct and ordered observed event times (censored and non-censored), $\mathcal{T} = \{t_i|t_i > t_{i-1} > \ldots > t_0\}$,

$$\ell_{\mathrm{cal}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) = \frac{1}{|\mathcal{T}|}\sum_{t_i \in \mathcal{T}}\left\|\hat{S}_{\mathrm{PKM}}^{p(t)}(t_i) - \hat{S}_{\mathrm{PKM}}^{g_\theta(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon})}(t_i)\right\|_1, \quad (11)$$

where $\hat{S}_{\mathrm{PKM}}$ is formulated as:

$$\hat{S}_{\mathrm{PKM}}(t_i) = \left(1 - \frac{\sum_{n:l_n=1}H(\hat{T}_n - t_{i-1}) - H(\hat{T}_n - t_i)}{M - \sum_{n=1}^{M}H(t_{i-1} - \hat{T}_n)}\right) \\ \times \hat{S}_{\mathrm{PKM}}(t_{i-1}), \quad (12)$$

and $H(b) = \frac{1}{2}(\mathrm{sign}(b) + 1)$ is the Heaviside step function. When evaluating the objective, $\ell_{\mathrm{cal}}(\boldsymbol{\theta}; \mathcal{D})$ in (11), $\hat{T}_n$ is either a sample from the model, $\hat{T}_n = g_\theta(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon})$, or an observed time $\hat{T}_n \sim p(t)$, for $\hat{S}_{\mathrm{PKM}}^{g_\theta(r_\psi(\boldsymbol{x}), \boldsymbol{\epsilon})}(t_i)$ or $\hat{S}_{\mathrm{PKM}}^{p(t)}(t_i)$, respectively. Expression $\hat{S}_{\mathrm{PKM}}$ represents the point-estimate-based formulation of the Kaplan Meier estimator, see [12] for details.

3.4 Learning

For joint learning of all model parameters, $\{\boldsymbol{c}_k\}_{k=1}^{K}$, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, we optimize both the latent representation and time-to-event (accuracy and calibration) objectives. The complete objective function for the proposed Survival Cluster Analysis (SCA) model is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^{K}; \mathcal{D}) = \ell_{\mathrm{dp}}(\boldsymbol{\psi}, \{\boldsymbol{c}_k\}_{k=1}^{K}; \mathcal{D}) \\ + \lambda_2\ell_{\mathrm{acc}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}) + \lambda_3\ell_{\mathrm{cal}}(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{D}), \quad (13)$$

where $\lambda_2, \lambda_3 > 0$ are hyper-parameters controlling the trade-off between accuracy and calibration objectives, relative to the clustering objective in (7). For simplicity and comparability with SFM, we set $\lambda_2 = \lambda_3 = 1$. The objective in (13) is optimized using stochastic gradient descent on minibatches from $\mathcal{D}$.

In practice, $\{\boldsymbol{c}_k\}$ is updated according to stochastic gradient descent by optimizing the KL objective (7), and is initialized with K-means after pretraining (13) without the clustering objective. During inference, we assign a new observation, $\boldsymbol{x}_\star$, to a cluster by first evaluating $q(u_\star = k|\boldsymbol{x}_\star)$ for $k = 1, \ldots, K$, then obtaining a hard assignment according to $u_\star = \mathrm{argmax}_k\, q(u_\star = k|\boldsymbol{x}_\star)$.

The maximum number of mixture components K is fixed during learning. However, provided that the KL divergence (7) encourages mixture proportions to follow a *stick-breaking process*, the *effective number of mixture components*, i.e., those with non-empty observation assignments, will be smaller than K, thus effectively resulting in the model learning the number of mixture components. This is illustrated in Figure 3 and described below in the experiments. The number of degrees of freedom, ν is a hyperparameter, set to 1 in our experiments.

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	FRAMINGHAM
Events (%)	23.9	27.5	68.1	51.0	23.8	11.14
N	394,823	7,894	9,105	68,082	5026	40,078
d (categorical)	729 (106)	26 (21)	59 (31)	789 (771)	206	12 (8)
Missing (%)	1.9	2.1	12.6	23.4	18.2	0.33
t_{max}	365	5,215	2,029	120	5,794	6,000

(a) Cluster assignments

(b) t-SNE plot

(c) Inferred minibatch K
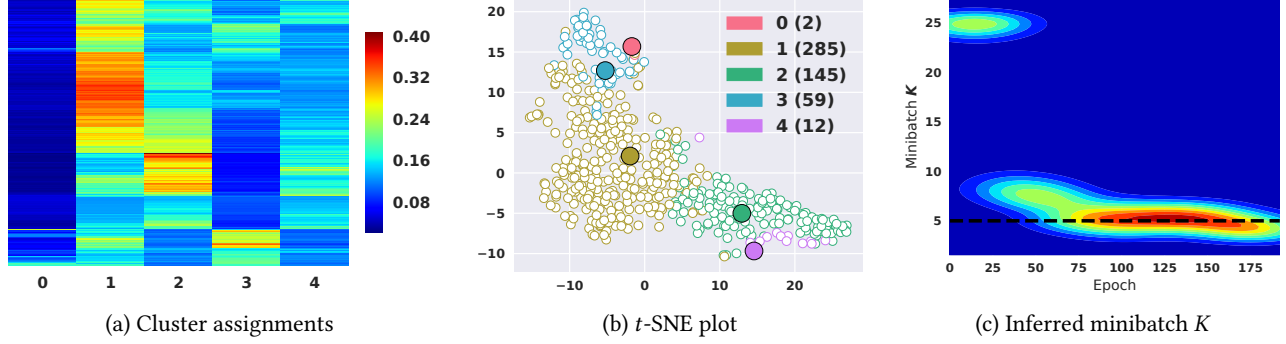
Figure 3: Inferred clusters on the testing set of SLEEP dataset, with $K = 25$ and $\gamma_o = 3$ where: (a) corresponding individual probability distribution $q(\pi|\xi, \{x_1\}_{n=1}^M)$, are approximated according to (5), (b) joint t-SNE plot of centroids c_k with latent representation z and (c) density plot of inferred number of clusters K during training.

4 EXPERIMENTS

The comparisons presented below are made across a diverse range of six datasets, as summarized in Table 1. Refer to the Supplementary Material for all details concerning the experimental setup. Throughout the experiments, we set $K = 25$ and select $\gamma_o = \{2, 3, 4, 8\}$ via grid search cross-validation from the training sets. TensorFlow code to replicate experiments can be found at https://github.com/paidamoyo/survival_cluster_analysis.

Datasets. Table 1 shows the summary statistics of the six datasets considered. The datasets are diverse in number of observations N, varying amounts of categorical (cat) and continuous covariates d, proportions of non-censored events, missingness rates in the $N \times d$ covariate matrix, and time horizon t_{max} (measured in days, except for SEER which is measured in months). Following information-theoretic data processing inequality conclusions from [37], demonstrating insignificant performance change relative to pre-imputation, we impute continuous and categorical covariates with the median and mode, respectively. In our experiments we do not convert time to a common scale but model it as is.

Publicly accessible datasets include: *i)* FLCHAIN: a study of non-clonal serum immunoglobin free light chains effects on survival time [15]. *ii)* SUPPORT: investigates the survival time of seriously-ill hospitalized adults [32]. *iii)* SEER: accessible from the Surveillance, Epidemiology, and End Results (SEER) Program. The dataset is preprocessed according to the details described in [44]. We restrict the dataset to a 10-year follow-up breast cancer subcohort.

The following datasets are available upon request: *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 [11]. *v)* SLEEP: a subset of the Sleep Heart Health Study (SHHS) [41], a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing. We focus on the baseline clinical visit and aggregated demographics, medications and questionnaire data as covariates. *vi)* FRAMINGHAM: a subset (Framingham Offspring) of the longitudinal study of heart disease [7] dataset, initially for predicting 10-year risk for future coronary heart disease (CHD).

Clustering Baselines. We consider the standard K-means and CoxPH based SSC-Bair [6] as strong clustering baselines for SCA. We provide quantitative evaluations in terms of the logrank score [36], and qualitative visualization of the clustering-based Kaplan-Meier sub-population survival curves.

Time-to-Event Baselines. We compare SCA to the following time-to-event baselines: SFM [12], DATE [11], S-CRPS [5], CoxPH [14], MTLR [51] and DRAFT [11]. From these, SFM and DATE are key to our comparisons because we leverage components from their formulation into SCA; namely, the accuracy loss from DATE and the distribution matching loss from SFM. In that sense, we expect SCA to perform as good as SFM and DATE, but with the added benefit of producing clusters with distinct risk profiles. We present quantitative evaluations in terms of C-index, Calibration slope, Relative Absolute Error (RAE), and mean Coefficient of Variation (CoV). Details of these metrics are provided in the Supplementary Material.

Table 2: Inferred cluster specific covariate information on the testing set for the FRAMINGHAM dataset. The inferred cluster assignments are according to the corresponding individual probability distribution $q(\pi|\xi, \{x_1\}_{n=1}^{M})$, approximated according to (5). Ranges in parentheses are 50% empirical ranges over (median) test-set predictions for the continuous and proportions for categorical covariates.

| Covariates	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Continous							
Age	$56_{(48,62)}$	$50_{(43,58)}$	$59_{(52,63)}$	$55_{(48,61)}$	$47_{(35,54)}$	$55_{(49,62)}$	$58_{(50,65)}$
HDL (mg/dL)	$43_{(37,53)}$	$52_{(44,63)}$	$67_{(59,85)}$	$54_{(45,66)}$	$62_{(55,70)}$	$41_{(35,48)}$	$42_{(36,52)}$
Total Cholesterol	$198_{(193,207)}$	$176_{(168,183)}$	$266_{(250,285)}$	$220_{(207,236)}$	$148_{(138,157)}$	$251_{(235,275)}$	$173_{(158,188)}$
Systolic Blood Pressure	$126_{(117,137)}$	$110_{(102,119)}$	$141_{(130,153)}$	$115_{(106,125)}$	$110_{(102,117)}$	$126_{(115,139)}$	$132_{(120,147)}$
Categorical							
Hypertension medication (Yes)	25.5%	4.97%	40.1%	11.3%	1.1%	41.6%	41.0%
Diabetic (Yes)	6.9%	2.63%	3.0%	3.3%	0.0%	20.8%	16.7%
Gender (Female)	36.9%	82.5%	63.6%	69.6%	74.5%	33.4%	36.4%
Current smoker (Yes)	23.9%	14.6%	28.0%	16.6%	22.3%	45.6%	25.1%
Race (Black)	16.4%	3.5%	29.5%	1.5%	8.5%	21.7%	27.7%
Race (Chinese)	4.2%	2.6%	0.0%	1.5%	2.1%	1.1%	3.0%
Race (Hispanic)	5.0%	2.3%	2.3%	2.3%	1.0%	4.0%	5.0%
Race (White)	74.4%	91.5%	68.2%	85.7%	88.3%	73.2%	64.2%
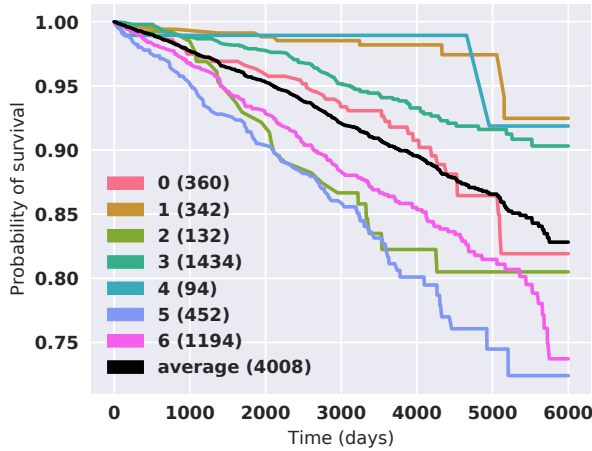
Figure 4: Inferred Cluster specific Kaplan-Meir Curves on the testing set of FRAMINGHAM dataset, with $K = 25$ and $\gamma_o = 8$. The inferred clusters assignment is according to the corresponding individual probability distribution $q(\pi|\xi, \{x_1\}_{n=1}^{M})$, approximated according to (5).

4.1 Qualitative Results

Figure 3 shows for the SLEEP dataset a) estimated individualized cluster assignment probability distributions (rows) evaluated according to (5); b) t-SNE plots of the model inferred centroids, c_k, as well as the individual latent representation $z = r_\psi(x)$; and c) density plot of the inferred number of (non-empty) clusters K during training. See the Supplementary Material for similar figures for all the other datasets, where we also include corresponding Kaplan-Meier curves, as in Figure 1.

Interestingly, the cluster-specific covariate statistics for the FRAMINGHAM dataset, which has the least number of covariates, are

shown in Table 2 and are consistent with findings from the Framingham Heart Study [7], which identified high blood cholesterol and high blood pressure as major risk factors for cardiovascular disease.

We obtain the cluster specific Kaplan-Meir curves illustrated in Figure 4 with corresponding cluster specific covariate information shown in Table 2. The inferred individual cluster assignment is obtained according to the individual probability distribution $q(\pi|\xi, \{x_1\}_{n=1}^{M})$, approximated according to (5). We consider curves above the population average low-risk while the curves below to be high-risk.

Therefore, our model identifies three high-risk clusters, indexed by $2, 5, 6$: i) cluster 6 and cluster 5 have similar statistics, as they both consists disproportionately of diabetic individuals on hypertension medication with elevated total cholesterol (normal is below 200), high systolic blood pressure (normal is below 120), and noticeably low HDL (normal is greater than 60); ii) cluster 2, is also driven by age, which is expected, where about 40% of the population is on hypertension medication, and with the worst systolic blood pressure and cholesterol compared to other clusters; iii) lower-risk clusters 1, 3, 4 are mostly comprised of females with normal levels of HDL, total cholesterol and systolic blood pressures; iv) cluster 0 represents the average statistics of the Framingham dataset, thus the survival curves directly follows the empirical population survival. Finally, note that the three high-risk clusters (2, 5 and 6) have a substantial over-representation of African Americans, known to have an increased risk for cardiovascular disease [7]. See the Supplementary Material for additional inferred cluster specific Kaplan-Meir curves on all datasets.

We demonstrate that by jointly learning clustering with respect to both the covariates x and predicted time-to-event t, our model SCA can identify high-, medium- and low-risk individuals , which is essential for clinical decision making. During inference, both the risk profile and individualized time-to-event can provide a comprehensive prediction mechanism for identifying cluster-based risk
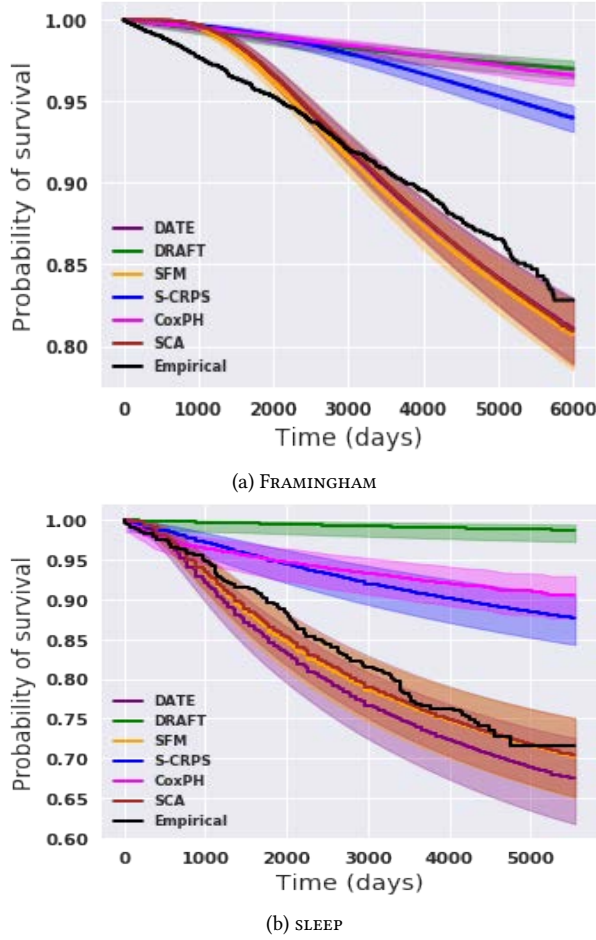
(a) FRAMINGHAM

(b) SLEEP

Figure 5: Survival function estimates for (a) FRAMINGHAM and (b) SLEEP data. Ground truth (Empirical) is compared to predictions from six models (DATE, DRAFT, SCA (our proposed model), SFM, S-CRPS and CoxPH). Error bars (shaded area) are calculated according to the Greenwood's formula [22].

factors, cluster-based risk profiles and individualized time predictions. Further, the advantage of matching the empirical mixture distribution with a (truncated) DP yields sparse predictions of cluster assignment probabilities, $q(\pi|\xi, \{x_1\}_{n=1}^{M})$, manifested as high confidence cluster assignments illustrated as a heatmap Figure 3(a).

Calibration Curves. We visually compare calibration curves from DATE, DRAFT, SCA, SFM, S-CRPS and CoxPH. Figure 5 shows the estimated *populations-based* model survival functions according to [12] and empirical Kaplan-Meier for the FRAMINGHAM and SLEEP datasets. Error bars (shaded area) are calculated according to the Greenwood's formula [22]. For all datasets, SCA- and SFM-estimated population survival functions closely match the empirical ground truth survival function, which is consistent with the high calibration slopes results in Table 3. See the Supplementary Material for additional calibration and survival function results on all datasets.

4.2 Quantitative Results

Below we describe performance metrics across all datasets and models. Specifically, calibration slope, mean CoV (coefficient of variation), C-index [23] and Relative Absolute Error [RAE, 51] provide a comprehensive evaluation, as they offer insights into consistency of time-to-event predictions, concentration of predicted distributions, pairwise ranking consistency, and accuracy of event time predictions, respectively. The results demonstrate that by jointly modeling the time-to-event and cluster assignments we obtain a better calibrated model, that is competitive in C-index, concentrated and accuracy of predictions. Table 3 shows the calibration slopes and RAE across all datasets and models. See the Supplementary Material for detailed RAE, mean CoV and C-index results.

Table 4 presents the clustering performances of the best performing K-means, SSC-Bair and SCA algorithms, measured in terms of the logrank score [36], for SSC-Bair and K-means we selected the best performing model from the set $K = \{2, 3, 4, 5, 6\}$.

Calibration Slope. For calibration we use the framework developed in SFM to evaluate the models [12]. An ideal calibration slope is 1, while a slope < 1 and slope > 1 indicates whether the model tends to underestimate or overestimate risk, respectively. The clustering objective in SCA augments the calibration objective we borrow from SFM, thus improving the calibration even for *non-iid* observations, such as FRAMINGHAM and EHR, which are considered poorly calibrated, as illustrated in SFM. Given that SCA leverages the calibration objective of SFM, it is not surprising that both SCA and SFM are competitive, followed by DATE, S-CRPS, MTLR, CoxPH and lastly DRAFT. See Supplementary Material for qualitative calibration plots.

Relative Absolute Error (RAE). We compute RAE for both censored and non-censored events. In Table 3 we present the RAE for non-censored event times ($l_n = 1$) for models that predict absolute event times, thus excluding scoring based models (CoxPH and MTLR). The results demonstrate that DATE, SFM and SCA (nonparametric) methods outperform DRAFT and S-CRPS (parametric) methods, which is expected since they all use a similar accuracy-aware objective function. For censored events ($l_n = 0$), RAE provides the lower bound error given the censored time provides tail information of $p(t|x)$; parametric methods (DRAFT and S-CRPS) have small advantage over nonparametric methods (SFM, DATE and SCA). See the Supplementary Material for additional results on censored event times.

Concordance Index (C-index). C-index is a ranking metric that does not account for uncertainty in time-to-event predictions. Therefore to evaluate the time-to-event models (except CoxPH) in terms of C-index, we use point summaries of the individualized time-to-event distributions, specifically, $\hat{t} = \text{median}(\{t_{ns}\}_{s=1}^{200})$, where t_{ns} is a sample from the trained model, $t_{ns} = g_\theta(r_\psi(x_n), \epsilon_s)$ on the test set. Apart from the small covariates, the very low event rate FRAMINGHAM dataset and the small high event rate SUPPORT dataset, none of the models have a clear advantage on the C-index metric. This is not surprising because C-index with very low event rate is heavily influenced by the censored observations. Note, for MTLR, although we can compute the C-index at prespecified thresholds, we are unable to compute a global C-index.

Table 3: Calibration slope and RAE metrics on test data.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
Calibration slope						
DATE	0.7537	0.9668	0.9068	0.9161	0.9454	0.7737
DRAFT	3.2138	5.4183	2.9640	2.0763	25.2855	5.7345
S-CRPS	1.6246	1.9662	1.1795	1.1613	2.5746	2.6114
CoxPH	2.5543	1.9116	1.3909	1.4358	3.8278	4.9945
MTLR	2.1957	1.9449	1.2017	1.2476	2.4792	5.4878
SFM	0.7734	0.9807	0.9405	**0.9540**	1.0235	0.7626
SCA (proposed)	**0.8006**	**0.9900**	**1.0086**	0.9290	**1.0223**	**0.8044**
RAE (non-censored)						
DATE	**0.6107**	0.5222	0.6691	0.5289	0.5224	0.5122
DRAFT	0.7099	0.6399	0.7109	0.6097	0.7465	0.6697
S-CRPS	0.7240	0.6378	**0.4851**	0.5323	0.7330	0.8369
CoxPH	-	-	-	-	-	-
MTLR	-	-	-	-	-	-
SFM	0.6146	**0.5111**	0.6398	0.5294	**0.5162**	**0.5074**
SCA (proposed)	0.6186	0.5134	0.6295	**0.5193**	0.5424	**0.5074**

Table 4: Logrank score and standard errors in parentheses. The best performing K-means and SSC-Bair models were selected from the set $K = \{2, 3, 4, 5, 6\}$.

	EHR	FLCHAIN	SUPPORT	SEER	SLEEP	FRAMINGHAM
SCA	-	**278.49 (0.0)**	496.50 (0.0)	**4803.59 (0.0)**	**63.74 (0.0)**	123.31 (0.0)
SSC-Bair	409.93 (0.0)	4.27 (0.37)	**1204**.14 (0.0)	4084.78 (0.0)	20.37 (0.0)	**125.87** (0.0)
K-means	417.00 (0.0)	5.33 (0.38)	99.06 (0.0)	3985.67 (0.0)	21.83 (0.0)	88.87 (0.0)

Coefficient of Variation (CoV) . Models that characterize the event time density function $f(t|\boldsymbol{x})$ result in uncertainty-aware time-to-event predictions. In practice, it is highly desirable for a model to generate concentrated time-to-event predictions. The CoV (coefficient of variation) measures the dispersion in a distribution; a Cov > 1 indicates high variance, while CoV < 1 indicates low variance distributions. Cov results provided in the Supplementary Material demonstrate that DATE, SCA and SFM are consistently low-variance distributions, followed by S-CRPS and lastly DRAFT. We cannot compute CoV for both MTLR and CoxPH. CoxPH estimates risk score, and therefore cannot be evaluated on CoV. MTLR does not specify the conditional hazards, $h(t|\boldsymbol{x})$, and thus we cannot recover $f(t|\boldsymbol{x}) = S(t|\boldsymbol{x})h(t|\boldsymbol{x})$.

Logrank Score. The logrank score is a nonparametric statistic that evaluates the similarity between a pair of survival functions, yielding high values for curves that are highly unlikely to be similar [36]. Further, the logrank statistic is especially powerful for measuring differences between survival functions that follow the Cox proportional hazard assumption, *i.e*, the survival functions do not cross. For K clusters, we compute $\binom{K}{2}$ pairwise comparisons. Table 4 demonstrates that our proposed SCA is the best performing method,

followed by SS-Bair and lastly K-means. Interestingly, SCA is unable to recover any clustering structure from the EHR dataset, as it is a homogeneous population of Type-2 diabetes subjects, whereas K-means and SSC-Bair are always able to produce clusters (which may be misleading for homogeneous datasets). This supports the need to account for survival information when clustering survival datasets, as both SCA and SSC-Bair incorporate time information in their clustering approaches.

5 CONCLUSIONS

We have developed the first time-to-event model for inferring individualized risk-based cluster assignments, while jointly predicting the time-to-event. Leveraging a Bayesian nonparametric stick-breaking representation of the Dirichlet Process, we have presented a method for learning a clustering structure in a latent representation, for which the number of clusters is unknown. We have demonstrated the need to account for time information when clustering survival datasets. Our model identifies interpretable and phenotyopically heterogeneous subpopulations, which are critical in a clinical setting for identifying subjects with diverse risk profiles. Extensive experiments demonstrate that the joint modeling approach yields substantial performance gains in calibration and

logrank scores, while remaining competitive in preserving pairwise ordering, predicting concentrated and accurate distributions. In the future, we plan to extend this work to account for locally-consistent, calibrated and accurate predictions within identified subpopulations.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Abramowitz, I. A. Stegun, and R. H. Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.
[2] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The lancet Diabetes & endocrinology, 2018.
[3] T. Ahmad, M. J. Pencina, P. J. Schulte, E. O'Brien, D. J. Whellan, I. L. Piña, D. W. Kitzman, K. L. Lee, C. M. O'Connor, and G. M. Felker. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. Journal of the American College of Cardiology, 2014.
[4] A. M. Alaa and M. van der Schaar. Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks. In NeurIPS, 2017.
[5] A. Avati, T. Duan, K. Jung, N. H. Shah, and A. Ng. Countdown regression: Sharp and calibrated survival predictions. arXiv, 2018.
[6] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. PLoS biology, 2(4):e108, 2004.
[7] E. J. Benjamin, D. Levy, S. M. Vaziri, R. B. D'agostino, A. J. Belanger, and P. A. Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. Jama, 1994.
[8] C. M. Bishop. Pattern recognition and machine learning. springer, 2006.
[9] D. M. Blei, M. I. Jordan, et al. Variational inference for dirichlet process mixtures. Bayesian analysis, 1(1):121–143, 2006.
[10] O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009.
[11] P. Chapfuwa, C. Tao, C. Li, C. Page, B. Goldstein, L. Carin, and R. Henao. Adversarial time-to-event modeling. In ICML, 2018.
[12] P. Chapfuwa, C. Tao, L. Carin, and R. Henao. Survival function matching for calibrated time-to-event predictions. arXiv, 2019.
[13] F. S. Collins and H. Varmus. A new initiative on precision medicine. New England journal of medicine, 2015.
[14] D. R. Cox. Regression models and life-tables. In Breakthroughs in statistics. Wiley Online Library, 1992.
[15] A. Dispenzieri, J. A. Katzmann, R. A. Kyle, D. R. Larson, T. M. Therneau, C. L. Colby, R. J. Clark, G. P. Mead, S. Kumar, L. J. Melton, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In Mayo Clinic Proceedings, 2012.
[16] U. Djuric, G. Zadeh, K. Aldape, and P. Diamandis. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ precision oncology, 2017.
[17] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25):14863–14868, 1998.
[18] T. S. Ferguson. A bayesian analysis of some nonparametric problems. The annals of statistics, 1973.
[19] T. Fernández, N. Rivera, and Y. W. Teh. Gaussian processes for survival analysis. In NeurIPS, 2016.
[20] S. Fotso. Deep neural networks for survival analysis based on a multi-task framework. arXiv, 2018.
[21] S. Gaynor and E. Bair. Identification of relevant subtypes via preweighted sparse clustering. Biostatistics, 2013.
[22] M. Greenwood et al. A report on the natural duration of cancer. A Report on the Natural Duration of Cancer., 1926.
[23] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modelling strategies for improved prognostic prediction. Statistics in medicine, 1984.
[24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.
[25] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association, 2001.

[26] H. Ishwaran and M. Lu. Random survival forests. Wiley StatsRef: Statistics Reference Online, 2008.
[27] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In IJCAI, 2017.
[28] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Machine learning, 1999.
[29] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology, 2018.
[30] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
[31] D. G. Kleinbaum and M. Klein. Survival analysis. Springer, 2010.
[32] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, et al. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. Annals of internal medicine, 1995.
[33] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In AAAI, 2018.
[34] C. Lee, W. R. Zame, A. M. Alaa, and M. van der Schaar. Temporal quilting for survival analysis. In AISTATS, 2019.
[35] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 2008.
[36] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep, 1966.
[37] X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In Machine Learning for Healthcare Conference, 2018.
[38] S. C. Mouli, L. Teixeira, B. Ribeiro, and J. Neville. Deep lifetime clustering. arXiv, 2019.
[39] E. Nalisnick and P. Smyth. Stick-breaking variational autoencoders. In ICLR, 2017.
[40] J. Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. Combinatorics, Probability and Computing, 2002.
[41] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, et al. The sleep heart health study: design, rationale, and methods. Sleep, 1997.
[42] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. Deep survival analysis. In Machine Learning for Healthcare Conference, 2016.
[43] C. E. Rasmussen. The infinite gaussian mixture model. In NeurIPS, 2000.
[44] L. A. G. Ries, J. L. Young Jr, G. E. Keel, M. P. Eisner, Y. D. Lin, and M.-J. D. Horner. Cancer survival among adults: US SEER program, 1988–2001. Patient and tumor characteristics SEER Survival Monograph Publication, 2007.
[45] J. Sethuraman. A constructive definition of dirichlet priors. Statistica sinica, 1994.
[46] S. J. Shah, D. H. Katz, S. Selvaraj, M. A. Burke, C. W. Yancy, M. Gheorghiade, R. O. Bonow, C.-C. Huang, and R. C. Deo. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation, 2015.
[47] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 2010.
[48] L. J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine, 1992.
[49] E. Xia, X. Du, J. Mei, W. Sun, S. Tong, Z. Kang, J. Sheng, J. Li, C. Ma, J. Dong, et al. Outcome-driven clustering of acute coronary syndrome patients using multi-task neural network with attention. arXiv preprint arXiv:1903.00197, 2019.
[50] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In ICML, 2016.
[51] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In NeurIPS, 2011.
[52] Q. Zhang and M. Zhou. Nonparametric Bayesian lomax delegate racing for survival analysis with competing risks. In NeurIPS, 2018.