Deep Survival Analysis: Nonparametrics and Missingness

Xenia Miscouridou XENIA.MISCOURIDOU@SPC.OX.AC.UK
University of Oxford
Oxford, Oxfordshire, OX1 3LB

Adler Perotte ADLER.PEROTTE@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

Noémie Elhadad NOEMIE.ELHADAD@COLUMBIA.EDU
Columbia University
New York City, NY, 10032

Rajesh Ranganath * RAJESHR@CIMS.NYU.EDU
New York University
New York City, NY 10003

Abstract

Clinical care requires understanding the time to medical events. Medical events include the time to a disease like chronic kidney disease progressing or the time to a complication as in stroke for high blood pressure. Models for event times live in the framework provided by survival analysis. We expand on *deep survival analysis* (Ranganath et al., 2016a), a deep generative model for survival analysis in the presence of missing data, where the survival times are modeled using Weibull distributions. We develop methods to relax the distributional assumptions in deep survival analysis using survival distributions that can approximate any true survival function. We show that the model structure mimics the information-optimal procedure in the presence of missing data. Our experiments demonstrate that moving to flexible survival functions yields better likelihoods and concordances for coronary heart disease prediction from electronic health records.

1. Introduction

Clinical Relevance. One of the core challenges to supporting clinicians in their care of patients is assessing risk of specific outcomes (e.g., onset of disease, deterioration of health status, death) in an automated, accurate, and meaningful fashion. Equipped with an estimated risk towards a particular event of interest, the clinician can better adjust care and prioritize treatments. In the field of machine learning in healthcare, there have been a variety of approaches proposed that directly estimate risk within specific pre-defined time frames. However, modeling the time to event of interest along with risk can provide a holistic picture from which to act upon.

* Corresponding author

Methods for building models of time-to-event data fall under survival analysis. At its core, survival analysis is probability distribution estimation. The probability distribution determines the survival function. The survival function at time t describes the fraction of observations that survive until that time. The goal is to provide a measure of prognosis based on measured covariates in medicine. This is provided by survival analysis through the survival probability distribution (Enker et al., 1995; D'amico et al., 1998; Pitt et al., 1999; Nath et al., 2004).

In survival analysis, there are censored observations, event times that are only known up to an interval. Censored observations commonly occur in medicine from patients leaving an institution or limited follow-up periods in experiments. The presence of censoring, distinguishes survival analysis and direct distribution estimation. In risk estimation for pre-defined time frames, survival provides the foundation for dealing with observations that left the study before the time frame was complete.

Traditional survival methods capture relationships between observed covariates, like age and gender, and the survival time. These methods typically require that the covariates are not missing and that there is a linear relationship between survival time and the covariate. Both of these requirements fail to be satisfied by large sources of clinical data such as electronic health records (EHRS). In EHRS, covariates like laboratory test values are fully observed only if they were required for the patient's care, which means most of them are missing (Hripcsak and Albers, 2012). Moreover, the relationship between covariates and survival times in healthcare can be nonlinear. For example, both high and low levels of blood pressure, increase mortality (Collaboration et al., 2002).

To address these issues Ranganath et al. (2016a) developed a deep latent variable model for survival analysis called deep survival analysis (DSA). While more flexible than traditional survival models, DSA makes distributional assumptions on the survival times. These assumptions place unneeded limits on the kinds of survival functions DSA represents. Further, while DSA tackles the case of missing data in the covariates, the structure of the model is cumbersome when the covariates are fully observed (not missing).

Technical Significance. In this work, we build upon the DSA model and address its limitations. We return to the view of survival analysis as a probability model specification with no missing covariates. We design desiderata based on distribution estimators that accurately model smooth densities, yet have tractable survival functions. We detail two families of distributions that satisfy the desiderata.

Next, we show that data imputation does not add information for predictions and that in the limit DSA models mimic the information-optimal procedure with missing data that builds a separate model for each Finally, we establish that a *limiting model* of DSA with flexible survival distributions can model any survival density. The limiting model corresponds to survival analysis where the missingness indicators are features and the missing values are filled in randomly. This limiting model can be fit directly when data are large enough. DSA provides a way to span a spectrum of survival models that trade off smoothness and representational power.

We study a large dataset of over 300,000 patient records built from the health records of a large metropolitan hospital to predict the time to coronary heart disease (CHD). We find that the added flexibility gives more accurate prediction of times to event.

The paper is structured as follows. Section 2 reviews survival analysis and deep survival analysis. Section 3 describes flexible survival estimation when all covariates are measured. Section 4 moves to the missing data case. Section 5 and 6 discuss results and future directions.

2. Deep Survival Analysis

The failure time distribution, in survival analysis models, measures the time to an event of interest. These times may be time to death, time to an organ failure, or time to developing a particular disease such as a heart disease. A survival model builds a distribution from covariates $\boldsymbol{x}$ to survival times $t: \; p(t \mid \boldsymbol{x})$. Example survival models include Cox proportional hazards (Cox, 1972) and Weibull regression (Ibrahim et al., 2005). These traditional approaches capture limited relationships between $\boldsymbol{x}$ and t and require the covariates to be fully observed. Deep survival analysis (DSA) proposes a survival analysis technique that models both the covariates and the survival times conditional on a shared latent process. The shared latent process $\boldsymbol{z}$ comes from a deep exponential family (DEF) (Ranganath et al., 2015) and couples the covariates and the survival times. Conditional on the latent process, DSA models the survival time via the Weibull distribution with scale controlled by the latent variable and model parameters $\boldsymbol{a}, \boldsymbol{b}$. Let k be the shape of the Weibull, then the generative process for DSA is

$$\begin{aligned} \boldsymbol{z_n} &\sim \text{DEF} \\ \boldsymbol{x}_n &\sim p(\cdot \mid \boldsymbol{\beta}, \boldsymbol{z_n}) \\ t_n &\sim \text{Weibull}\left(\log(1+\exp\left(\boldsymbol{z}_n^\top \boldsymbol{a} + \boldsymbol{b}\right), k\right). \end{aligned} \tag{1}$$

DSA handles missing covariates that are prevalent in electronic health records without having to impute them because missing $\boldsymbol{x}$ can be integrated out. Given covariates, the model predicts failure times using the predictive distribution

$$p(t \mid \boldsymbol{x}) = \int_z p(t \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{x}) d\boldsymbol{z}.$$

The Weibull assumption on $p(t \mid \boldsymbol{z})$ places limitations on the predictive distribution $p(t|\boldsymbol{x})$ no matter the prior on $p(\boldsymbol{z})$. To see this, let $\mathbb{H}$ denote the entropy. The entropy relates to the spread of the distribution. Its conditional variant, conditional entropy, relates to the average spread of each conditional distribution. Then by the conditional entropy inequality (Cover and Thomas, 2012),

$$\mathbb{H}(t \mid \boldsymbol{x}) \geq \mathbb{H}(t \mid \boldsymbol{z}, \boldsymbol{x}) = \mathbb{H}(t \mid \boldsymbol{z}).$$

This means the predictive distribution must have higher entropy than the Weibull in $p(t \mid \boldsymbol{z})$ regardless of the choice of $p(\boldsymbol{z})$. This limitation stems from the choice of Weibull distribution with fixed shape in DSA. In the next section, we develop flexible survival likelihoods.

3. Nonparametric survival likelihoods

For survival data $(t, \boldsymbol{x})$, all survival analysis methods specify a distribution for t given any $\boldsymbol{x}$. A survival model defines a class of $p(t \mid \boldsymbol{x})$. In this section we describe the desiderata on the survival

likelihoods that model $p(t \mid \boldsymbol{x})$ in the case of no missing data and no latent structure $\boldsymbol{z}$. However, the same distributions can be used to model $p(t \mid \boldsymbol{z})$, under the latent variable model and the presence of missing data.

To build a survival model, we lay out two objectives:

- The model family should be flexible enough to approximate any smooth distribution.
- Computing the survival function should be straightforward.

The first objective ensures that the model family does not introduce error in the large data limit. The second ensures the likelihood of censored data in the presence of independent censoring is simple to compute. With independent censoring, the likelihood of a right-censored[1] observation is the survival function. We develop two examples that meet these two desiderata. The first builds on invertible transforms, while the second uses a discrete approximation.

Survival distribution via invertible transforms. A deterministic transformation changes one probability distribution into another. We use this to show that the true failure distribution $d_{\boldsymbol{x}}$ can be represented as a transformation of another distribution. The following proposition formalizes this.

Proposition 1 *Let $d_{\boldsymbol{x}}$ denote the true failure time distribution for each covariate $\boldsymbol{x}$ that is continuous with respect to Lebesgue measure on $\mathbb{R}_0^+$. Let $s_{\boldsymbol{x}}$ be a density function of a distribution that is absolutely continuous on $\mathbb{R}_0^+$. Then, there exists an invertible transformation function g_x such that*

$$\begin{aligned} \tilde{t} &\sim s_{\boldsymbol{x}}(\cdot) \\ g_{\boldsymbol{x}}\left(\tilde{t}\right) &\stackrel{d}{=} t \sim d_{\boldsymbol{x}}(\cdot). \end{aligned} \tag{2}$$

Proof If $d_{\boldsymbol{x}}$ and $s_{\boldsymbol{x}}$ are continuous, their cumulative distribution functions (CDFs) denoted by $D_{\boldsymbol{x}}$ and $S_{\boldsymbol{x}}$ respectively are invertible. By the probability integral transform theorem, we know $S_{\boldsymbol{x}}\left(\tilde{t}\right)$ is uniform and $D_{\boldsymbol{x}}^{-1}$ of a uniform random variable has density function $d_{\boldsymbol{x}}$. Therefore, setting $g_{\boldsymbol{x}}\left(\tilde{t}\right) = D_{\boldsymbol{x}}^{-1}\left(S_{\boldsymbol{x}}\left(\tilde{t}\right)\right)$ shows that $g_{\boldsymbol{x}}\left(\tilde{t}\right) \stackrel{d}{=} t$. ■

The result shows that there exists an invertible transformation that can map from any smooth distribution to another. This means that if we can construct flexible invertible mappings, we can approximate arbitrary densities. Satisfying the second desideratum of easy-to-compute survival functions requires an additional constraint. Let $g_{\boldsymbol{x}}$ be the transformation of a distribution with density $s_{\boldsymbol{x}}$, then the CDF of the transform at time T is

$$\int_{g_{\boldsymbol{x}}^{-1}(0)}^{g_{\boldsymbol{x}}^{-1}(T)} s_{\boldsymbol{x}_i}(g_{\boldsymbol{x}}^{-1}(t))dt = S_x\left(g_{\boldsymbol{x}}^{-1}(T)\right) - S_x\left(g_{\boldsymbol{x}}^{-1}(0)\right).$$

This means the CDF of the transformed variable is easy to compute if the original CDF is easy-to-compute. This is satisfied by Weibull distributions, log-Normal distributions, and exponential distributions (Ross, 2014).

1. All that is known for right censored observation is that the true survival time is greater or equal to the censoring time.

One way to construct a flexible family of invertible functions is to define it directly, say via monotonic transformations. Another way to build a rich family of functions is to compose simpler ones. Consider, a sequence of L invertible transformations $g^1, \ldots, g^L$, then the function composition $g = g^L \circ \cdots \circ g^1$ is also invertible. We model the random time coming from the true distribution as a transformation from an initial distribution $s_{\boldsymbol{x}}$ parametrized by $\boldsymbol{W}$,

$$\begin{aligned} t_i^0 &\sim s_{\boldsymbol{x}_i}(t_i; \boldsymbol{W}_t) \\ t_i^\ell &= h_\ell(t_i^{\ell-1}; \boldsymbol{\theta}^\ell, \boldsymbol{x}), \\ &\cdots \\ t_i^L &= h_L(t_i^{L-1}; \boldsymbol{\theta}^L, \boldsymbol{x}) \end{aligned}$$

The above is a type of normalizing flow (Rezende and Mohamed, 2015). Other examples of transformations are linear transformations in the log-space of t_i, planar and radial flows, and any other invertible parametric family. For certain classes of these functions, the inverse may exist—but require an iterative process to compute. To ameliorate this computational cost, we can use flows defined from their inverse functions (Ranganath et al., 2016b).

Survival distribution via discrete/categorical distributions. We present a flexible discrete distribution for survival times. This discrete distribution can be used both when interested in a discrete time distribution (for example time being the number of months a patient will survive) and continuous distributions.

Consider a discrete bounded time domain with $T < \infty$ number of bins. The model posits that a survival time falls in time interval k with probability given by the function $f_k(\boldsymbol{x}_i; \boldsymbol{W}_k)$.

$$t_i \sim \text{Categorical}\left(f_0\left(\boldsymbol{x}_i; \boldsymbol{W}_0\right), \ldots, f_T\left(\boldsymbol{x}_i; \boldsymbol{W}_T\right)\right) \tag{3}$$

The advantage of using the categorical distribution is that it can approximate arbitrary conditional distributions in the discrete domain and also approximate arbitrary smooth densities. This concept is formalized by the the following proposition.

Proposition 2 *Let F denote the true CDF of a differentiable distribution, $\hat{f}_n(\cdot)$ be the density estimator of a sample of size n with h the bin width. Then $\hat{f}_n$ is consistent as $h \to 0$ and $nh \to \infty$.*

The proof of this proposition follows from standard histogram consistency arguments (Wasserman, 2006). This discrete distribution falls into the transformation family of Proposition 1 if we relax the transformation to be measurable rather than invertible. This shows the first desideratum of flexible model families. The second desideratum of tractable survival functions follows because the CDF of the discrete distribution is a sum over a bounded domain.

The categorical approximation loses the continuity information that is implicit in healthcare settings; precisely, the probability of surviving t months is similar to the probability of surviving $t+1$ and $t-1$ months. This prior knowledge can be modeled by placing a smooth autoregressive prior on the parameters $\boldsymbol{W}_t$. An example of a prior is a linear dynamical system. With $\mathcal{N}$ denoting the normal distribution and σ^2 the variance parameter, the model for the prior is:

$$\boldsymbol{W}_t \sim \mathcal{N}(\boldsymbol{W}_{t-1}, \sigma^2).$$

If we infer the parameters $\boldsymbol{W}_t$ via maximum-a-posteriori, this prior acts as a regularizer and forces the parameters and the probabilities to be close.

Estimation. To estimate the latent variable model in Equation (1) with the aforementioned nonparametric likelihoods, we can use black box variational inference methods (Ranganath et al., 2014; Kingma and Welling, 2014; Rezende et al., 2014). Variational inference casts inference to optimization over a variational family. We specify the family to be in the same class as the generative model.

4. Nonparametric survival likelihoods and missing data

Having developed survival models for fully observed $\boldsymbol{x}$, we now turn to the case of missing covariates. A common approach to handling missing covariates builds models to predict the missing values in a process called imputation (Rubin, 1996). Imputation approaches range from filling in mean values across the observed population to building complex nonlinear models. Imputation shines when the scientific questions involve covariate values themselves. For example, what is the effect of sodium on blood pressure? However, imputation provides no extra information for joint predictions.

To demonstrate that imputation adds no information more formally, let $\boldsymbol{\beta}$ be a parameter of an imputation model learned on observed training covariates $\boldsymbol{x}_{\text{train}}^{\text{observed}}$, missingness indicators $\boldsymbol{b}_{\text{train}}$, and survival times t_{train}. The parameters $\boldsymbol{\beta}$ are learning from the training data $(\boldsymbol{x}_{\text{train}}^{\text{observed}}, \boldsymbol{b}_{\text{train}}, t_{\text{train}})$, thereby making them a (stochastic) function of the training data. Now consider a test point with covariates and missingness indicators $\boldsymbol{x}^{\text{observed}}, \boldsymbol{b}$ and survival time t. Using the imputation model learned over the training data, we have the data processing chain,

$$t \to (\boldsymbol{x}^{\text{observed}}, \boldsymbol{b}) \xrightarrow{\boldsymbol{\beta}} (\boldsymbol{x}^{\text{observed}}, \boldsymbol{b}, \boldsymbol{x}^{\text{imputed}}) \to (\boldsymbol{x}^{\text{observed}}, \boldsymbol{x}^{\text{imputed}}),$$

that imputes missing data for a single test example. If I is the mutual information, by the data processing inequality we have:

$$I[t; \boldsymbol{x}^{\text{observed}}, \boldsymbol{b}] \geq I[t; \boldsymbol{x}^{\text{observed}}, \boldsymbol{x}^{\text{imputed}}]. \tag{4}$$

This inequality demonstrates that for each fixed imputation model parameter $\boldsymbol{\beta}$, the data completed by imputation has the same or less information with the survival time than the observed data and missingness indicators. Since averages maintain inequalities, the inequality Equation (4) holds when averaging over training datasets or uncertainty in the imputation model parameters.

We now show that limiting variants of DSA models approximate building models for each pattern of missingness. The latent variables in DSA induce dependencies between the survival times and covariates:

$$p(t \mid \boldsymbol{x}) = \int p(t \mid \boldsymbol{z})\, p(\boldsymbol{z} \mid \boldsymbol{x}) d\boldsymbol{z}.$$

This equation shows that to construct a DSA model that can predict directly from the observed data and the missingness, the latent variable must encode the observed data values and missingness indicators. We start by letting $\boldsymbol{z}$ have twice the dimensionality of the $\boldsymbol{x}$. For simplicity, let the data be real valued with Lebesgue absolutely continuous density. Let $\mathcal{N}$ be the normal distribution, and define $\mathbb{I}$ as

the indicator function, which takes value one when its argument is true. Then, given a prior and likelihood with parameter π and a fixed value c,

$$\begin{aligned} z &= (z_1, z_2) \\ p(z_1) &\sim \mathcal{N}(0, 1) \\ p(z_2) &\sim \text{Bernoulli}(\pi) \\ p(\boldsymbol{x} \mid \boldsymbol{z}) &= \mathbb{I}[\boldsymbol{x} = z_1]\mathbb{I}[z_2 = 1] + \mathbb{I}[\boldsymbol{x} = c]\mathbb{I}[z_2 = 0]. \end{aligned}$$

When dimension d is observed, this forces the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ to be

$$\begin{aligned} p(\boldsymbol{z}_{1,d} \mid \boldsymbol{x}) &= \boldsymbol{x}_d \\ p(\boldsymbol{z}_{2,d} \mid \boldsymbol{x}) &= 1, \end{aligned}$$

since the probability that $\boldsymbol{x}_d$ equals c is zero by Lebesgue absolute continuity. When dimension d is missing

$$\begin{aligned} p(\boldsymbol{z}_{1,d} \mid \boldsymbol{x}) &= \mathcal{N}(0, 1) \\ p(\boldsymbol{z}_{2,d} \mid \boldsymbol{x}) &= \text{Bernoulli}(\pi). \end{aligned}$$

Let superscript missing denote the missing dimensions and observed denote the observed dimensions, then the predictive distribution is

$$\begin{aligned} p(t \mid \boldsymbol{x}) &= \int p(t \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{x}) d\boldsymbol{z} \\ &= \int p(t \mid \boldsymbol{x}^{\text{observed}}, 1^{\text{observed}}, \boldsymbol{z}_1^{\text{missing}}, \boldsymbol{z}_2^{\text{missing}}) p(\boldsymbol{z}_1^{\text{missing}}, \boldsymbol{z}_2^{\text{missing}}) d\boldsymbol{z}_1^{\text{missing}} \boldsymbol{z}_2^{\text{missing}}. \end{aligned}$$

For small π, $\boldsymbol{z}_2^{\text{missing}}$ is zero with high probability. This means as π gets small, the latent variable can encode both the missing data and what was measured. This encoding result can be generalized to covariates over a discrete space and continuous variables with a finite number of discrete masses by adding in a dummy value disjoint from the discrete masses. Combined with the flexible likelihoods from the previous section, DSA can approximate arbitrary survival functions.

We call the model with π tending to zero, the *limiting model*. This limiting model corresponds to building a density estimate with the missingness indicators and the unobserved covariates set to a random value or dummy value:

$$\log p(t \mid \boldsymbol{x}^{\text{observed}}, \boldsymbol{b}, \boldsymbol{x}^{\text{missing}} = c). \tag{5}$$

Since nothing is missing in the conditioning set, we can estimate Equation (5) without latent variables using maximum likelihood. To parametrize the limiting model we can use the flexible survival likelihoods defined in the previous section such as the categorical one in Equation (3). One tradeoff between the limiting approach and the latent variable approach lies in the ease at which relationships between variables and the survival time can be extracted. In DSA with latent variables, computing $p(\boldsymbol{z} \mid x)$ along with Monte Carlo reveals the relationship between x and t, while the limiting model requires marginalizing over missingness indicators.

5. Evaluation of Survival Models

To assess the value of the different survival models, we turn to concordance index (CI) and log likelihood as two complementary evaluation measures. Survival models are traditionally evaluated on the CI. CI on a test set measures the fraction of pairs of test examples whose predicted failure times are ordered correctly with respect to their true failure times. This means concordance does not measure the accuracy of an individual failure time, rather it only measures the relative ordering of times (rankings) amongst patients. To illustrate this fact with a concrete example, suppose one model predicts survival times with perfect accuracy, then take another model that outputs ten times the survival times of the previous model. Both of these models have the same concordance because the second model preserves the ordering of survival times, but with incorrect survival times.

Alternatively, likelihood-based evaluations prefer to match the predicted time to the actual time. That is, when the model assigns all of its probability mass to the true time, the likelihood is maximal. In this sense, if survival times are of interest, likelihoods are a better choice than concordance. However, it is noteworthy that two models can produce similar likelihoods and yet yield different CI. For instance, consider a model that incorrectly predicts the time to event for a patient that had a very early outcome. In that case, the CI will be widely affected because that patient is now outranked mistakenly by many other patients. While for the likelihood, this mistake affects only one patient.

The survival distributions we describe are nonparametric. This means if there is enough data, then the error in the models go to zero regardless of the evaluation metrics chosen. However, if the capacity of the models are limited, choosing a likelihood or a different divergence for the task at hand may help. For example, proportional hazards (Cox, 1972) relates to ranking and thus concordance (Steck et al., 2008) and modeling log-failure times with a log-Normal distribution relates to evaluations based on getting orders of magnitude of the prediction correct. Maximum likelihood corresponds to KL-divergence minimization, alternatively divergences such as the Jensen-Shannon divergence can be developed for estimation. However, censoring makes direct estimation more challenging. Since survival functions are built on probability estimates, proper scoring rules (Gneiting and Raftery, 2007) can be adapted for survival estimation (Avati et al., 2018).

6. Experiments

We study our nonparametric likelihoods and limiting estimators on survival times for coronary heart disease (CHD). CHD is also known as ischemic heart disease or coronary artery disease. CHD is a significant cause of mortality and of expensive, chronic conditions like congestive heart failure.

6.1. Data

The dataset includes information of 300K individuals from a large metropolitan hospital. The population comprises of adults with at least 5 interactions with the hospital's network. The data covariates we use are 9 vital signs, 80 laboratory test measurements, 5K medication orders, and 13K diagnosis codes. All the data was aggregated at a month level giving observations per patient for any month that the patient had an observed measurement. Data with real values were averaged out for each month whereas categorical data were encoded as binary variables. Patients appear in the record

a variable amount of times. We sample patient-month pairs inversely proportional to the number of times a patient interacted with the hospital. CHD events were defined by the occurrence of any ICD-9 diagnosis code with the following prefixes: 413 (angina pectoris), 410 (myocardial infarction), or 411 (coronary insufficiency).

6.2. Baseline and Model Setup

We compare against the original DSA model from Ranganath et al. (2016a) at its best complexity (K=50). For our approach, we study three models: (i) the categorical survival likelihood with latent variables with support up to 400 months, (ii) the limiting categorical survival likelihood with support up to 400 months, and (iii) an invertible transformation of a Weibull by a tanh and linear layer using the limiting model. All comparisons were performed with $K = 50$. For inference in the latent variable models, we use black box variational inference. All non-linear transformations were three layer ReLU functions such as for the f_t in the categorical likelihoods. We use a batch size of 1,000 and parallelize within batches. For optimization, we use RMSProp (Tieleman and Hinton, 2012) with Nesterov momentum. For diagnosis and medications, we use Bernoulli likelihoods that allow for fast evaluation when most of the observations are zero. Let $\beta^{\text{meds}}_{\boldsymbol{W}_i}$ and $\beta^{\text{meds}}_{b_i}$ denote weight and bias parameters, then the likelihood of the ith medication in the nth observation is

$$p(x^{\text{meds}}_{n,i} \mid \beta^{\text{meds}}_{\boldsymbol{W}_i}, \beta^{\text{meds}}_{b_i}, z_n) \sim \text{Bernoulli}(1 - \exp((z_n^\top \beta^{\text{meds}}_{\boldsymbol{W}_i} + \beta^{\text{meds}}_{b_i})^2)),$$

This likelihood allows for both positive and negative correlations between medication observations and can be computed in time proportional to the number of taken medications. It generalizes the BerPo link functions developed in Zhou (2015) which are typically limited to nonnegative $\boldsymbol{z}$ and $\boldsymbol{W}$. For the laboratory tests and vitals, we use the Student-T likelihood from Ranganath et al. (2016a), where the parameters of the Student-T come from a multilayer transformation of the latent variable $\boldsymbol{z}$. Given a multilayer transformation f with parameters $\beta^{\text{labs}}_{W_i}$ and degrees of freedom ν, the likelihood of the ith lab in the nth observation is

$$p(x^{\text{labs}}_{n,i} \mid \beta^{\text{labs}}_{W_i}, z_n) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}} \left(1 + \frac{(x^{\text{labs}}_{n,i} - (f_{\beta^{\text{labs}}_{W_i}}(z_n))^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

The Student-T likelihood reduces the influence of irregularities commonly present in clinical data (Ghassemi et al., 2018).

6.3. Evaluation

We split the data randomly with 25,000 patients for validation and 25,000 for testing. The validation set is used to choose the best model based on the evaluation criteria. We evaluate based on the concordance index and predictive likelihood. The concordance checks whether failure times are ordered correctly. As orderings can only be checked on pairs where at least one of the items failed, concordance does not incur as great a cost for uncensored observations.

6.4. Results

We find that all the new proposed models outperform the baseline DSA according to their concordance. We find the unweighted categorical models (Categorical and Limiting Categorical in Table 1) perform best, both according to likelihood and CI. Note that the likelihoods were computed as an average of predictive likelihoods over one million patient months and concordances were computed over a billion pairs.

Method	Predictive Likelihood	Concordance index
DSA	-0.45	0.70
Categorical	-0.42	0.72
Limiting Categorical	**-0.41**	**0.73**
Limiting Transformed Weibull	-0.46	0.73

Table 1: Predictive likelihood and concordance of deep survival analysis for different models. We find that the categorical models perform best overall, but all of the new models based on flexible survival likelihoods outperform the baseline on concordance.

7. Discussion

We built off DSA, a multilayer approach for survival analysis. We show that the survival likelihood choices in DSA restrict the survival distributions DSA can represent. We construct nonparametric likelihoods for survival analysis that can be used when the covariates are completely observed. We show that imputation adds no information and show that DSA model can approximate the information-optimal procedure for missing data. We use the flexbile likelihoods to develop a more accurate survival model for coronary heart disease. Many future directions remain to pursue, for example building new flexible transforms that are appropriate for electronic health records, and developing regularization techniques for the limiting variant such that there is smoothness between different sets of missing values.

References

Anand Avati, Tony Duan, Kenneth Jung, Nigam H Shah, and Andrew Ng. Countdown regression: Sharp and calibrated survival predictions. *arXiv preprint arXiv:1806.08324*, 2018.

Prospective Studies Collaboration et al. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet*, 360(9349):1903–1913, 2002.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

David Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):87–22, 1972.

Anthony V D'amico, Richard Whittington, S Bruce Malkowicz, Delray Schultz, Kenneth Blank, Gregory A Broderick, John E Tomaszewski, Andrew A Renshaw, Irving Kaplan, Clair J Beard, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Jama*, 280(11):969–974, 1998.

Warren E Enker, Howard T Thaler, Milicent L Cranor, and Tatyana Polyak. Total mesorectal excision in the operative treatment of carcinoma of the rectum. *Journal of the American College of Surgeons*, 181(4):335–346, 1995.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

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

George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.

Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian survival analysis*. Wiley Online Library, 2005.

Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings ICLR (International Conference on Learning Representations)*, 2014.

Jayant Nath, Elyse Foster, and Paul A Heidenreich. Impact of tricuspid regurgitation on long-term survival. *Journal of the American College of Cardiology*, 43(3):405–409, 2004.

Bertram Pitt, Faiez Zannad, Willem J Remme, Robert Cody, Alain Castaigne, Alfonso Perez, Jolie Palensky, and Janet Wittes. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *New England Journal of Medicine*, 341(10):709–717, 1999.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114, 2016a.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016b.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Sheldon M Ross. *Introduction to probability and statistics for engineers and scientists*. Academic Press, 2014.

Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

Harald Steck, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216, 2008.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Larry Alan Wasserman. *All of nonparametric statistics*. Springer, 2006.

Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *Artificial Intelligence and Statistics*, pages 1135–1143, 2015.