Neural Topic Models with Survival Supervision: Jointly Predicting Time-to-Event Outcomes and Learning How Clinical Features Relate

Linhong Li[1][0000−0001−6274−1371], Ren Zuo[2][0000−0002−6160−4081],
Amanda Coston[1][0000−0001−9282−9921], Jeremy C. Weiss[1][0000−0003−1693−9082],
and George H. Chen[1,✉][0000−0001−8645−051X]

Carnegie Mellon University, Pittsburgh PA 15213, United States
[1]{linhongl,acoston,jweiss2,georgech}@andrew.cmu.edu,
[2]renzuo.wren@gmail.com

Abstract. In time-to-event prediction problems, a standard approach to estimating an interpretable model is to use Cox proportional hazards, where features are selected based on lasso regularization or stepwise regression. However, these Cox-based models do not learn how different features relate. As an alternative, we present an interpretable neural network approach to jointly learn a survival model to predict time-to-event outcomes while simultaneously learning how features relate in terms of a topic model. In particular, we model each subject as a distribution over "topics", which are learned from clinical features as to help predict a time-to-event outcome. From a technical standpoint, we extend existing neural topic modeling approaches to also minimize a survival analysis loss function. We study the effectiveness of this approach on seven healthcare datasets on predicting time until death as well as hospital ICU length of stay, where we find that neural survival-supervised topic models achieves competitive accuracy with existing approaches while yielding interpretable clinical "topics" that explain feature relationships.

Keywords: Survival Analysis · Topic Modeling · Interpretability

1 Introduction

Predicting the amount of time until a critical event occurs—such as death, disease relapse, or hospital discharge—is a central focus in the field of survival analysis. Especially with the increasing availability of electronic health records, survival analysis data in healthcare often have both a large number of subjects and a large number of features measured per subject. In coming up with an interpretable survival analysis model to predict time-to-event outcomes, a standard approach is to use Cox proportional hazards [6], with features selected using lasso regularization [25] or stepwise regression [12]. However, these Cox-based models do not inherently learn how features relate.

To simultaneously address the two objectives of learning a survival model for time-to-event prediction and learning how features relate specifically through

a topic model, Dawson and Kendziorski [8] combine latent Dirichlet allocation (LDA) [3] with Cox proportional hazards to obtain a method they call survLDA. The idea is to represent each subject as a distribution over topics, and each topic as a distribution over which feature values appear. The Cox model is given the subjects' distributions over topics as input rather than the subjects' raw feature vectors. Importantly, the topic and survival models are jointly learned.

In this paper, we propose a general framework for deriving neural survival-supervised topic models that is substantially more flexible than survLDA. Specifically, survLDA estimates model parameters via variational inference update equations derived specifically for LDA combined with Cox proportional hazards; to use another other sort of combination would require re-deriving the inference algorithm. In contrast, our approach combines any topic model and any survival model that can be cast in a neural net framework; combining LDA with Cox proportional hazards is only one special case. Importantly, our framework yields survival-supervised topic models that are interpretable so long as the underlying topic and survival models are interpretable. As a byproduct of taking a neural net approach, we can readily leverage many deep learning advances. For example, we can avoid deriving a special inference algorithm and instead use any neural net optimizer such as Adam [17] to learn the joint model in mini-batches, which scales to large datasets unlike survLDA's variational inference algorithm.

As numerous combinations of topic/survival models are possible, for ease of exposition, we demonstrate how to combine LDA with Cox proportional hazards in a neural net framework, yielding a neural variant of survLDA. We refer to our neural variant as survScholar since we build on scholar [5], a neural net approach to learning LDA and various other topic models. We benchmark survScholar on seven datasets, finding that it can yield performance competitive with various baselines while also yielding interpretable topics that reveal feature relationships. For example, on a cancer dataset, survScholar learns two topics that are associated with longer survival time, and one topic associated with lower survival time. The first two pro-survival topics provide different explanations for patients attributes correlated with surviving longer: one topic is associated with normal vital signs and laboratory measurements, while the other includes vital sign and laboratory derangements of sodium and creatinine. survScholar can help discover such feature relationships that clinicians could then verify. Meanwhile, when survScholar's prediction is inaccurate, examining the topics learned could help with model debugging.

2   Background

We begin with some background and notation on topic modeling and survival analysis. For ease of exposition, we phrase notation in terms of predicting time until death; other critical events are possible aside from death.

We assume that we have access to a training dataset of n subjects. For each subject, we know how many times each of d "words" appears, where the dictionary of words is pre-specified (continuous clinical feature values are discretized

into bins). As an example, one word might correspond to "white blood count reading in the bottom quintile"; for a given subject, we can count how many such readings the subject has had recorded in the past. We denote $X_{i,u}$ to be the number of times word $u \in \{1, \ldots, d\}$ appears for subject $i \in \{1, \ldots, n\}$. Viewing X as an n-by-d matrix, the i-th row of X (denoted by X_i) can be thought of as the feature vector for the i-th subject.

As for the training label for the i-th subject, we have two recordings: event indicator $\delta_i \in \{0, 1\}$ specifies whether death occurred for the i-th subject, and observed time $Y_i \in \mathbb{R}_+$ is the i-th subject's "survival time" (time until death) if $\delta_i = 1$ or the "censoring time" if $\delta_i = 0$. The idea is that when we stop collecting training data, some subjects are still alive. The i-th subject still being alive corresponds to $\delta_i = 0$ with a true survival time that is unknown ("censored"); instead, we know that the subject's survival time is at least the censoring time.

Topic Modeling A topic model transforms the i-th subject's feature vector X_i into a topic weight vector $W_i \in \mathbb{R}^k$, where $W_{i,g}$ is the fraction that the i-th subject belongs to topic $g = 1, 2, \ldots, k$. The $W_{i,g}$ terms are nonnegative and $\sum_{g=1}^{k} W_{i,g} = 1$. For example, LDA models topic weight vectors W_i's to be generated i.i.d. from a user-specified k-dimensional Dirichlet distribution. Next, to relate feature vector X_i with its topic weight vector W_i, let $\overline{X}_{i,u}$ denote the fraction of times a word appears for a specific subject, meaning that $\overline{X}_{i,u} = X_{i,u} / \left(\sum_{v=1}^{d} X_{i,v} \right)$. Then LDA assumes the factorization

$$\overline{X}_{i,u} = \sum_{g=1}^{k} W_{i,g} A_{g,u} \tag{2.1}$$

for a topic-word matrix $A \in \mathbb{R}^{k \times d}$. Each row of A is a distribution over the d vocabulary words and is assumed to be sampled i.i.d. from a user-specified d-dimensional Dirichlet distribution. In matrix notation, $\overline{X} = WA$. Standard LDA is unsupervised and, given matrix $\overline{X}$, estimates the matrices W and A.

Survival Analysis Standard topic modeling approaches like LDA do not solve a prediction task. To predict time-to-event outcomes, we turn toward survival analysis models. Suppose we take the i-th subject's feature vector to be $W_i \in \mathbb{R}^k$ instead of X_i. As this notation suggests, when we combine topic and survival models, W_i corresponds to the i-th subject's topic weight vector; this strategy for combining topic and survival models was first done by Dawson and Kendziorski [8], who worked off of the original supervised LDA formulation by McAuliffe and Blei [23] (which is not stated for survival analysis). We treat the training data as $(W_1, Y_1, \delta_1), \ldots, (W_n, Y_n, \delta_n)$, disregarding the "raw" feature vectors X_i's.

The standard survival analysis prediction task can be stated as using the training data $(W_1, Y_1, \delta_1), \ldots, (W_n, Y_n, \delta_n)$ to estimate, for any test subject with feature vector $w \in \mathbb{R}^k$, the subject-specific survival function

$S(t|w) = \mathbb{P}(\text{subject survives beyond time } t \mid \text{subject's feature vector is } w).$

Importantly, unlike standard regression where, for any test feature vector w, we predict a single real number, here we predict a whole function $S(\cdot|w)$.

Our neural survival-supervised topic modeling framework crucially requires that the we can construct a predictor $\widehat{S}(\cdot|w)$ for the subject-specific survival function $S(\cdot|w)$ by minimizing a differentiable loss. Numerous survival models satisfy this criterion. For example, consider the classical Cox proportional hazards model [6]. We learn a parameter vector $\beta \in \mathbb{R}^k$ that weights the features, i.e., prediction for an arbitrary feature vector $w \in \mathbb{R}^k$ is based on the inner product $\beta^\top w$. The differentiable loss function for the Cox model is

$$L_{\mathsf{Cox}}(\beta) = \sum_{i=1}^{n} \delta_i \Big[-\beta^\top W_i + \log \sum_{j=1 \text{ s.t. } Y_j \geq Y_i}^{n} \exp(\beta^\top W_j) \Big]. \qquad (2.2)$$

After computing parameter estimate $\widehat{\beta}$ by minimizing $L_{\mathsf{Cox}}(\beta)$, we can estimate survival functions $S(\cdot|w)$ via the following approach by Breslow [4]. Denote the unique times of death in the training data by $t_1, t_2, \ldots, t_m$. Let d_i be the number of deaths at time t_i. We first compute the so-called hazard function $\widehat{h}_i := d_i/(\sum_{j=1 \text{ s.t. } Y_j \geq Y_i}^{n} e^{\widehat{\beta}^\top W_j})$ at each time index $i = 1, 2, \ldots, m$. Next, we form the "baseline" survival function $\widehat{S}_0(t) := \exp(-\sum_{i=1 \text{ s.t. } t_i \leq t}^{m} \widehat{h}_i)$. Finally, subject-specific survival functions are estimated to be powers of the baseline survival function: $\widehat{S}(t|w) := [\widehat{S}_0(t)]^{\exp(\widehat{\beta}^\top w)}$.

3   Neural Survival-Supervised Topic Models

We now present our proposed neural survival-supervised topic modeling framework. Our framework can use any topic model that has a neural net formulation (e.g., neural versions of LDA [3], SAGE [10], and correlated topic models [20] are provided by Card et al. [5]; recent topic models like the Embedded Topic Model [9] can also be used). Moreover, our framework can use any survival model learnable by minimizing a differentiable loss (e.g., Cox proportional hazards [6] and its lasso/elastic-net-regularized variants [25], the Weibull accelerated failure time (AFT) model [15], and all neural survival models we are aware of). For ease of exposition, we focus on combining LDA with the Cox proportional hazards model, similar to what is done by Dawson and Kendziorski [8] except we do this combination in a neural net framework.

We first need a neural net formulation of LDA. We can use the SCHOLAR framework by Card et al. [5]. Card et al. do not explicitly consider survival analysis in their setup although they mention that predicting different kinds of real-valued outputs can be incorporated by using different label networks. We use their same setup and have the final label network perform survival analysis. We give an overview of SCHOLAR before explaining our choice of label network.

The SCHOLAR framework specifies a generative model for the data, including how each individual word in each subject is generated. In particular, recall that $X_{i,u}$ denotes the number of times the word $u \in \{1, 2, \ldots, d\}$ appears for the

i-th subject. Let v_i denote the number of words for the i-th subject, i.e., $v_i = \sum_{u=1}^{d} X_{i,u}$. We now define the random variable $\psi_{i,\ell} \in \{1, 2, \ldots, d\}$ to be what the ℓ-th word for the i-th subject is (for $i = 1, 2, \ldots, n$ and $\ell = 1, 2, \ldots, v_i$). Then the generative process for SCHOLAR with k topics is as follows, stated for the i-th subject:

1. Generate the i-th subject's topic distribution:
   (a) Sample $\widetilde{W_i}$ from a logistic normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$.
   (b) Set the topic weights vector for the i-th subject to be $W_i = \text{softmax}(\widetilde{W_i})$.
2. Generate the i-th subject's words:
   (a) Set word parameter $\phi_i = f_{\text{word}}(W_i)$, where f_{word} is a generator network.
   (b) For word $\ell = 1, 2, \ldots, v_i$: Sample $\psi_{i,\ell} \sim \text{Multinomial}(\text{softmax}(\phi_i))$.
3. Generate the i-th subject's output label:
   Sample Y_i from a distribution parameterized by label network $f_{\text{label}}(W_i)$.

Different choices for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, f_{word}, and f_{label} lead to different topic models. The parameters are learned via amortized variational inference [18, 24]. To approximate LDA where topic distributions are sampled from a symmetric Dirichlet distribution with parameter $\alpha > 0$, we set $\boldsymbol{\mu}$ to be the all zeros vector, $\boldsymbol{\Sigma} = \text{diag}((r-1)/(\alpha r))$, and $f_{\text{word}}(w) = w^\top H$ where $H \in \mathbb{R}^{k \times d}$ has a Dirichlet prior per row. We describe how to set f_{label} to obtain survival supervision next.

Survival Supervision To incorporate the Cox survival loss, we change step 3 of the generative process above to be deterministic and output the variable $\Xi_i = f_{\text{label}}(W_i) := \beta^\top W_i$ for parameter vector $\beta \in \mathbb{R}^k$. In particular, we do not model how observed times Y_i's are generated; modeling Ξ_i's is sufficient. Then we can minimize the Cox proportional hazards loss from equation (2.2), rewritten to use the variables Ξ_i's that are parameterized by β:

$$L_{\text{Cox}}(\beta) = \sum_{i=1}^{n} \delta_i \Big[-\Xi_i + \log \sum_{j=1 \text{ s.t. } Y_j \geq Y_i}^{n} \exp(\Xi_i) \Big], \text{ where } \Xi_i = \beta^\top W_i. \quad (3.1)$$

For a hyperparameter $\eta > 0$ that weights the importance of the survival loss, the final overall loss that gets minimized is the sum of $\eta L_{\text{Cox}}(\beta)$ and SCHOLAR's topic model loss (given by the negation of equation (4) in the SCHOLAR paper [5]). We refer to the resulting model as survSCHOLAR.

We remark that rewriting the Cox loss to use Ξ_i variables (for which we can replace the inner product $\Xi_i = \beta^\top W_i$ with a neural net $\Xi_i = g(W_i)$) is by Katzman et al. [16] and also works for the Weibull AFT model.

Model Interpretation For the g-th topic learned, we can look at its distribution over words $A_g \in \mathbb{R}^d$ (given in equation (2.1)) and, for instance, rank words by their probability of appearing for topic g (our experiments later rank words using a notion of comparing to background word frequencies). The g-th topic is also associated with Cox regression coefficient β_g, where $\beta = (\beta_1, \beta_2, \ldots, \beta_k) \in \mathbb{R}^k$ is the parameter from equation (3.1). Under the Cox model, β_g being larger means that the g-th topic is associated with *shorter* survival times.

Table 1. Basic characteristics of the survival datasets used.

Dataset	Description	# subjects	# features	% censored
SUPPORT-1	acute resp. failure/multiple organ sys. failure	4194	14	35.6%
SUPPORT-2	COPD/congestive heart failure/cirrhosis	2804	14	38.8%
SUPPORT-3	cancer	1340	13	11.3%
SUPPORT-4	coma	591	14	18.6%
UNOS	heart transplant	62644	49	50.2%
METABRIC	breast cancer	1981	24	55.2%
MIMIC(ICH)	intracerebral hemorrhage	1010	1157	0%

4   Experimental Results

Data We conduct experiments on seven datasets: data on severely ill hospitalized patients from the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) [19], which—as suggested by Harrell [11]—we split into four datasets corresponding to different disease groups (acute respiratory failure/multiple organ system failure, cancer, coma, COPD/congestive heart failure/cirrhosis); data from patients who received heart transplants in the United Network for Organ Sharing (UNOS);[1] data from breast cancer patients (METABRIC) [7]; and lastly patients with intracerebral hemorrhage (ICH) from the MIMIC-III electronic heath records dataset [14]. For all except the last dataset, we predict time until death; for the ICH patients, we predict time until discharge from a hospital ICU. Basic characteristics of these datasets are reported in Table 1. We randomly divide each dataset into a 80%/20% train/test split. Our code is available and includes data preprocessing details.[2]

Experimental Setup We benchmark SURVSCHOLAR against a total of 7 baselines: 4 classical methods (Cox proportional hazards [6], lasso-regularized Cox [25], k-nearest neighbor Kaplan-Meier [2, 22], and random survival forests (RSF) [13]), 2 deep learning methods (DeepSurv [16] and DeepHit [21]), and a naive two-stage decoupled LDA/Cox model (fit unsupervised LDA first and then fit a Cox model). For all methods, 5-fold cross-validation on training data is used to select hyperparameters (if there are any) prior to training on the complete training data. Hyperparameter search grids are included in our code. For both cross-validation and evaluating test set accuracy, we use the time-dependent concordance C^{td} index [1], which roughly speaking is the fraction of pairs of subjects in a validation or test set who are correctly ordered, accounting for temporal and censoring aspects of survival data. Similar to area under the ROC curve for classification, a C^{td} index of 0.5 corresponds to random guessing and 1 is a perfect score. For every test set C^{td} index reported, we also compute its 95% confidence interval, which we obtain by taking bootstrap samples of the test set with replacement, recomputing the C^{td} index per bootstrap sample, and taking the 2.5 and 97.5 percentile values.

---

[1]We use the UNOS Standard Transplant and Analysis Research data from the Organ Procurement and Transplantation Network as of September 2019, requested at: https://www.unos.org/data/

[2]https://github.com/lilinhonglexie/NPSurvival2020

Table 2. Test set C^{td} indices with 95% bootstrap confidence intervals.

Model	Dataset						
	SUPPORT-1	SUPPORT-2	SUPPORT-3	SUPPORT-4	UNOS	METABRIC	MIMIC(ICH)
COX	0.630	0.571	**0.569**	0.592	0.583	0.664	0.610
	(0.606, 0.655)	(0.538, 0.604)	(0.531, 0.607)	(0.537, 0.649)	(0.575, 0.592)	(0.622, 0.706)	(0.564, 0.652)
LASSO-COX	0.627	0.567	0.556	0.603	0.557	0.664	**0.667**
	(0.604, 0.652)	(0.535, 0.600)	(0.517, 0.594)	(0.538, 0.666)	(0.548, 0.565)	(0.623, 0.708)	(0.621, 0.712)
K-NN	0.601	0.581	0.557	0.501	0.584	0.669	0.563
	(0.577, 0.628)	(0.545, 0.614)	(0.517, 0.592)	(0.432, 0.576)	(0.576, 0.592)	(0.627, 0.708)	(0.518, 0.612)
RSF	0.602	**0.604**	0.568	0.492	0.587	**0.697**	0.651
	(0.575, 0.628)	(0.570, 0.636)	(0.530, 0.601)	(0.414, 0.575)	(0.579, 0.595)	(0.659, 0.736)	(0.602, 0.697)
DEEPSURV	0.636	0.555	0.555	0.602	0.580	0.686	0.616
	(0.611, 0.660)	(0.521, 0.589)	(0.517, 0.591)	(0.548, 0.659)	(0.572, 0.589)	(0.644, 0.725)	(0.571, 0.661)
DEEPHIT	0.633	0.579	0.547	0.590	**0.598**	0.683	0.598
	(0.607, 0.660)	(0.548, 0.609)	(0.511, 0.585)	(0.518, 0.657)	(0.590, 0.606)	(0.644, 0.721)	(0.553, 0.649)
NAIVE LDA/COX	0.586	0.565	0.525	**0.607**	0.537	0.661	0.599
	(0.559, 0.611)	(0.533, 0.595)	(0.486, 0.563)	(0.541, 0.672)	(0.528, 0.545)	(0.622, 0.698)	(0.549, 0.646)
SURVSCHOLAR	0.630	0.587	0.568	0.567	0.588	0.690	0.619
	(0.604, 0.655)	(0.553, 0.618)	(0.528, 0.605)	(0.509, 0.625)	(0.580, 0.595)	(0.649, 0.731)	(0.572, 0.661)
SURVSCHOLAR-FEW	**0.637**	0.580	0.568	0.586	0.588	0.695	0.590
	(0.612, 0.662)	(0.547, 0.610)	(0.528, 0.605)	(0.532, 0.640)	(0.581, 0.596)	(0.656, 0.735)	(0.547, 0.632)

For SURVSCHOLAR, we also include a variant SURVSCHOLAR-FEW that instead of picking whichever hyperparameters (number of topics k and the survival loss importance weight η) achieve the highest training cross-validation C^{td} index, we instead favor choosing a hyperparameter setting with the fewest number of topics that achieves a cross-validation C^{td} index within 0.005 of the best score. We empirically found that often a much fewer number of topics achieves a training cross-validation score that is nearly as good as the max found. For ease of model interpretation, using a fewer number of topics is preferable.

Results Test set C^{td} indices are reported in Table 2 with 95% confidence intervals. The main takeaways are that: (a) the two SURVSCHOLAR variants are the best or nearly the best performers on SUPPORT-1, SUPPORT-3, and METABRIC; (b) even when the SURVSCHOLAR variants are not among the best performers, they still do as well as some established baselines; (c) the two SURVSCHOLAR variants have very similar performance (so for interpretation, we use SURVSCHOLAR-FEW), and (d) no single method is the best across all datasets.

Next, we interpret the learned topic models. We plot the topics learned by SURVSCHOLAR-FEW for the SUPPORT-3 dataset on cancer patients in Fig. 1: each topic is a column in the plot, where above each topic, we denote its Cox β regression coefficient (higher means shorter survival time); rows correspond to features. Deeper red colors indicate features that occur more for a topic; color intensity values are multiplicative ratios compared to background word frequencies and are explained in more detail in the appendix. The three topics in this SUPPORT-3 cancer dataset indicate one anti-survival and two pro-survival topics. There is a primary anti-survival topic described by old age, multicomorbidity, hyponatremia, and hyperventilation. The first pro-survival topic describes vital sign and laboratory derangements including hypernatremia, elevated creatinine, hypertension, and hypotension. The second pro-survival topic with slightly stronger pro-survival association suggests otherwise-healthy patients with normal vital signs and laboratory measurements.

We summarize our findings for the other datasets. For SUPPORT-1, SUPPORT-2, SUPPORT-4, UNOS, and METABRIC, only two topics (corresponding to healthy and unhealthy) are identified per dataset by SURVSCHOLAR-FEW. For the MIMIC(ICH) dataset, SURVSCHOLAR-FEW has similar prediction performance as deep learning baseline DeepHit (c.f., Table 2) but neither method performs as well as lasso-regularized Cox. By inspecting the 5 topics learned by SURVSCHOLAR-FEW, we find the topics difficult to interpret as too many features are surfaced as highly probable. In this high-dimensional setting where the number of features is larger than the number of subjects, we suspect that regularizing the model (e.g., by replacing LDA with SAGE [10]) is essential to obtaining interpretable topics. Our interpretations of learned topic models for all datasets along with additional visualizations are available in our code repository.
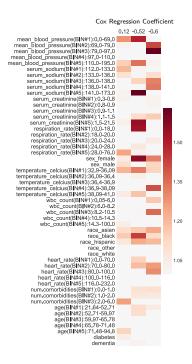

Fig. 1. Topics learned for SUPPORT-3. Rows index features, columns index topics.

5   Discussion

Despite many methodological advances in survival analysis with the help of deep learning, these advances have mostly not focused on interpretability. Model interpretation can be especially challenging when there are many features and how they relate is unknown. In this paper, we show that neural survival-supervised topic models provide a promising avenue for learning structure over features in terms of "topics" that help predict time-to-event outcomes. These topics can be used by practitioners to check if learned topics agree with domain knowledge and, if not, to help with model debugging. Rigorous evaluations of other neural survival-supervised topic models aside from fusing LDA with Cox are needed to better understand which combinations of topic and survival models yield both highly accurate time-to-event predictions and clinically interpretable topics.

Bibliography

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

[2] R. Beran. Nonparametric regression with randomly censored survival data. *Technical report, University of California, Berkeley*, 1981.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2013.

[4] N. Breslow. Discussion of the paper by D. R. Cox (1972) cited below. *Journal of the Royal Statistical Society, Series B*, 34(2):216–217, 1972.

[5] D. Card, C. Tan, and N. A. Smith. Neural models for documents with metadata. In *Proceedings of Association for Computational Linguistics*, 2018.

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

[7] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, and Y. Yuan. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

[8] J. A. Dawson and C. Kendziorski. Survival-supervised latent dirichlet allocation models for genomic analysis of time-to-event outcomes. *arXiv preprint arXiv:1202.5999*, 2012.

[9] A. B. Dieng, F. J. Ruiz, and D. M. Blei. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*, 2019.

[10] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. In *International Conference on Machine Learning*, pages 1041–1048, 2011.

[11] F. E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2015.

[12] 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*, 3(2):143–152, 1984.

[13] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[14] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[15] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd ed. edition, 2002.

[16] 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*, 18(1):24, 2018.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[19] 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, and N. Desbiens. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.

[20] J. D. Lafferty and D. M. Blei. Correlated topic models. In *Advances in Neural Information Processing Systems*, pages 147–154, 2006.

[21] 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 Conference on Artificial Intelligence*, 2018.

[22] D. J. Lowsky, Y. Ding, D. K. Lee, C. E. McCulloch, L. F. Ross, J. R. Thistlethwaite, and S. A. Zenios. A K-nearest neighbors survival probability prediction method. *Statistics in Medicine*, 32(12):2062–2069, 2013.

[23] J. D. McAuliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

[24] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[25] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 2011.

A   Interpreting Topic Heatmaps

In this appendix, we explain how to interpret our topic heatmaps (Fig. 1 and additional plots in our code repository). For many topic models including LDA, a topic is represented as a distribution over d vocabulary words. SCHOLAR [5] (and also our survival-supervised version SURVSCHOLAR) reparameterizes these topic distributions; borrowing from SAGE [10], SCHOLAR represents a topic as a deviation from a background log-frequency vector. This vector accommodates common words that have similar frequencies across data points. When we visualize a topic, we take this modeling approach into account and only choose to highlight features that have positive log-deviations from the background. Given a topic, having positive log-deviation is analogous to having higher conditional probabilities in the classic topic modeling case but explicitly is relative to background word frequencies (rather than being raw topic word probabilities).

To fill in the details, in step 2(a) of SURVSCHOLAR's generative process (stated in Section 3), each word is drawn from the conditional distribution softmax$(\gamma + w^T B)$, where $\gamma \in \mathbb{R}^d$ is the background log-frequency vector, $w \in \mathbb{R}^k$ contains a sample's topic membership weights, and $B \in \mathbb{R}^{k \times d}$ encodes (per topic) every vocabulary word's log-deviation from the word's background. This is a reparameterization of how LDA is encoded, which has each word drawn from the conditional distribution softmax$(w^T H)$ for $H \in \mathbb{R}^{k \times d}$. In particular, note that $H_g = \gamma + B_g$ for every topic $g \in \{1, 2, \ldots, k\}$. The background log-frequency vector γ is learned during neural net training. Note that SAGE [10] further encourages sparsity in B by adding ℓ_1 regularization on B.

We found ranking words within a topic by their raw probabilities (A_g in equation (2.1)) to be less interpretable than ranking words based on their deviations from their background frequencies (B_g) precisely because commonly occurring background words make interpretation difficult. In fact, when Dawson and Kendziorski [8] introduced SURVLDA, they used an ad hoc pre-processing step to identify background words to exclude from analysis altogether. We avoid this pre-processing and use log-deviations from background frequencies instead.

In heatmaps such as the one in Fig. 1, each column corresponds to a topic. For the g-th topic, instead of plotting its raw log-deviations (encoded in $B_g \in \mathbb{R}^d$), which are harder to interpret, we exponentiated each word's log-deviation to get the word's multiplicative ratio from its background frequency (i.e., we compute $\exp(B_g)$); the color bar intensity values are precisely these multiplicative ratios of how often a word appears relative to the word's background frequency.

To highlight features that distinguish topics from one another, we also sort rows in the heatmap by descending differences between the largest and smallest values in a row. Thus, features whose deviations vary greatly across topics tend to show up on the top. A technical detail is that we sorted with respect to the original features, rather than the one-hot encoded or binned features. Therefore, as an example, all bins under mean blood pressure stay together. For features associated with multiple rows in the heatmap, we computed the difference between the largest and smallest values for each row, and used the largest difference (across rows) for sorting.