Uncertainty Estimation in Cancer Survival Prediction

Anonymous authors
Paper under double-blind review

Abstract

Survival models are used in various fields, such as the development of cancer treatment protocols. Although many statistical and machine learning models have been proposed to achieve accurate survival predictions, little attention has been paid to obtain well-calibrated uncertainty estimates associated with each prediction. The currently popular models are opaque and untrustworthy in that they often express high confidence even on those test cases that are not similar to the training samples and even if their predictions are wrong. We propose a Bayesian framework for survival models that not only gives more accurate survival predictions but also quantifies the survival uncertainty better. Our approach is a novel combination of variational inference for uncertainty estimation, neural multi-task logistic regression for estimating nonlinear and time-varying risk models, and an additional sparsity-inducing prior to work with high dimensional data.

1 Introduction

Survival modeling is key to precision oncology wherein cancer management, and treatment planning is personalized to a patient's clinical, pathological, demographical, and genomic state. Aided by the digitization of medical records, several studies over the past four decades have collected survival data based on longitudinal follow-up for various patient cohorts. Modeling survival in a cohort based on covariates known at the time of prediction is a complex task because the covariates to be taken into account can be large in number. These covariates can be entangled with each other by their interdependencies and interactions. A good survival model should give both (a) accurate survival estimates, and (b) a well-calibrated measure of uncertainty. We address the second problem in this work, which has escaped the attention it deserves. Most of the machine learning models trade complexity and accuracy for interpretability, and may not indicate a drop in confidence even when their prediction is off on the test inputs, or the test inputs are outliers with respect to the training input data distribution.

Cox proportional hazards model (Cox-PH) proposed by Cox (1972) is one of the oldest and most popular statistical models to predict survival. Cox-PH uses a hazard function to model the survival in a cohort and assumes that a patient's relative log-risk of treatment failure (disease recurrence or death) at any time is a linear combination of the patient's covariates that scales the underlying hazard function, which is another restrictive assumption. Multi-task logistic regression (MTLR) was proposed by Yu et al. (2011) as a remedy to the assumption of temporal constancy of relative risk between two patients, which led to increased prediction accuracy over Cox-PH. MTLR uses multi-task regression and joint likelihood minimization to model log-risk in a given time interval as a linear combination of the covariates. Recently, neural-MTLR was proposed by Fotso (2018) to move away from the linearity assumption as well to increase the prediction accuracy. Neural-MTLR models nonlinear interactions among covariates as features extracted by the lower layers of a neural network, whose last layers are same as that of MTLR.

The above mentioned survival models are unable to access per-patient uncertainty in survival predictions. Uncertainty calibration is important if survival prediction models are to be deployed in clinical settings. The prediction of any model is usually untrustworthy when the test data from a new patient is out of the training distribution (OOD). In such OOD cases, it is important to involve human experts, and hence it is important to identify such cases with the model rightfully expressing high uncertainty or low confidence. Bayesian neural networks (BNNs) provide a framework to capture the underlying uncertainties inherent to both the data (data uncertainty) and the limitations of the model (model uncertainty). We propose a Bayesian extension of MTLR and neural-MTLR that can capture patient-specific survival uncertainties. We also show how capturing the uncertainty in the prediction helps us handle heterogeneous data and analyze prognostically important covariates. We further incorporate a sparsity inducing prior in our model to handle a large number of input covariates.

2 Theoretical Background and Proposed Method

2.1 Survival models

The setting that we assume has a set of covariates x_i associated with each patient i, a time to adverse event (T_i) (usually death or disease recurrence), and an event indicator (E_i). The event indicator $E_i = 1$ means that the patient died after a time interval of T_i. Patients with $E_i = 0$ are called right-censored, indicating that she was surviving (or living disease-free) at time T_i, but survival beyond that time in unknown.

The survival function and hazard function are two important outcomes of the survival models. The survival function, $S(t) = P(T \geq t)$, is the probability of a patient to survive more than time t. The hazard function $\lambda(t)$ is given by $\lim_{\Delta t \to 0} P(t \geq T \geq t + \Delta t \mid T \geq t)/\Delta t$, which means the probability that an individual will not survive an extra infinitesimal amount of time Δt, given they have already survived up to time t. Cox-PH (Cox, 1972) models the hazard function $\lambda(t|\vec{x})$ at time t for a given vector input covariates $\vec{x}$ in terms of an underlying hazard function $\lambda_0(t)$ and linear weights θ for the covariates as follows: $\lambda(t|\vec{x}) = \lambda_0(t)\exp(\vec{x}^T\vec{\theta})$.

MTLR (Yu et al., 2011) assumes a series of logistic regression models for $m+1$ time intervals, where m is chosen based on the desired fineness of temporal variation and the size of the training data, as follows: $P_{\theta_i}(T \geq t_i \mid x) = (1 + exp(\vec{\theta_i}.\vec{x} + b_i))^{-1}; 0 \leq i \leq m$. The parameters $\vec{\theta_i}$ and b_i depend on the time interval i, whereas the input vector $\vec{x}$ is same for all the regression models. However, the outputs of these logistic regression models are not independent, because a death event at time t_i would mean a death event at all subsequent time points $t_j, j > i$. We encode the output of the regression model, using a m-dimensional binary sequence $y = (y_1, y_2, y_3 ... y_m)$, where, $y_i = 0$ means that the patient is living at time t_i and $y_i = 1$ means that the patient is dead at time t_i. Thus, once we encounter a $y_i = 1$, all subsequent $y_j, j > i$ are bound to be 1. A smoothness prior on the parameters across time ensures that the predictions are not noisy. The probability of observing a sequence $y = (y_1, y_2, y_3 ... y_m)$ is the likelihood of the model. It can be generalized by the logistic regression model as follows: $P_\theta(Y = (y_1, y_2, y_3 ... y_m) \mid \vec{x}) = \left[\sum_{k=j}^m y_i(\vec{\theta}_i.\vec{x} + b_i)\right] / [\sum_{k=0}^m exp(f_\theta(\vec{x}, k))]$, where $f_\theta(\vec{x}, k) = \sum_{j=k+1}^m (\vec{\theta}_i.\vec{x} + b_i)$.

The loss function for uncensored patients is obtained by taking the logarithm of the joint likelihood term and adding regularization terms for temporal smoothness of the parameters and the resultant predictions, as follows:

$$L = \frac{C_1}{2} \sum_{j=1}^{m} \|\vec{\theta}_j\|^2 + \frac{C_2}{2} \sum_{j=1}^{m} \|\vec{\theta}_{j+1} - \vec{\theta}_j\|^2 - \sum_{i=1}^{n} \left[\sum_{j=1}^{m} y_j(s_i)(\theta_j.\vec{x} + b_j) - \log(\sum_{k=0}^{m} exp(f_\theta(\vec{x}_i, k)) \right] \tag{1}$$

where C_1 and C_2 are hyperparameters which control the amount of smoothing in the parameters and n is the number of patients.

For right-censored patients (those who are lost to follow-up), there are more than one consistent binary sequences of y_i's. In this case the likelihood of the patient is the sum of likelihoods of all possible sequences. The overall likelihood for censored patients whose last contact was closest to time point t_j is given as follows: $P_{\theta_i}(T \geq t_i \mid x) = \left[\sum_{k=j}^m exp(f_\theta(\vec{x}, k))\right] / [\sum_{k=0}^m exp(f_\theta(\vec{x}, k))]$.

Neural-MTLR (Fotso, 2018) models nonlinear combinations of the covariates as inputs to the MTLR model, where both the MTLR model and the nonlinear feature extraction are trained end-to-end using backpropagation (gradient descent) on a loss function similar to Equation 1.

2.2 Variational Inference

A feed-forward neural network trained with gradient descent will arrive at point estimates. However, in the case of Bayesian NNs (BNNs), the weights are not point estimates but a parameterized probability distribution. Our task is to find a distribution over the parameters given the input data, i.e., $p(\theta \mid D)$. With this posterior, we can predict test output y^* for a new test input x^* by marginalizing the likelihood over the parameters θ. However, even for the modest-sized NNs, the number of parameters prohibits an analytical calculation of uncertainty, and one has to resort to approximate inference methods. We define an approximating variational distribution $q_\psi(\theta)$ with parameters ψ. Then the Kullback-Leibler divergence (KL) with respect to the parameters ψ is minimized between the proposed posterior and the true posterior, as follows:

$$\mathrm{KL}(q_\psi(\theta)||p(\theta|D)) = \int q_\psi(\theta) \log \frac{q_\psi(\theta)}{p(\theta|D)} d\theta \tag{2}$$

Minimizing the KL divergence is equivalent to minimizing the variational free energy (Friston et al., 2007), (Blundell et al., 2015), where the latter is often computed on M mini batches $D^1, D^2, \ldots, D^M$ for computational tractability. We then estimate the cost using an unbiased Monte Carlo (MC) approximation for each mini batch as follows: $\theta^j \sim q_\psi(\theta)$, $\hat{\mathcal{L}}(\psi) = -\sum_{i \in D^j} \log(p(y_i|f^{\theta^j}(x_i))) + (1/M)\mathrm{KL}(q_\psi(\theta)||p(\theta))$.

2.3 Proposed probabilistic weights to model uncertainty

We assume the posterior and the prior on weights to be a spike and slab, which is standard for sparse linear models (Mitchell & Beauchamp, 1988) (George & McCulloch, 1993) (Titsias & Lázaro-Gredilla, 2011). Recently, a closed form expression for the KL divergence between the spike and slab posterior and spike and slab prior was derived by Tonolini et al. (2019), which we utilized in this work. The prior probability density is given as follows: $p_\psi(\theta) = \prod_{i=1}^{N}(\alpha\mathcal{N}(\theta_i; 0, 1) + (1-\alpha)\delta(\theta_i))$, where $\delta(.)$ is the dirac delta function centered at zero. The sparsity of solution can be increased for this prior by decreasing α from one to zero. The posterior is chosen to be of similar form, given as: $q_\psi(\theta) = \prod_{i=1}^{N}(\gamma_i\mathcal{N}(\theta_i; \mu_i, \sigma_i^2) + (1-\gamma_i)\delta(\theta_i))$, where μ_i, σ_i and γ_i are the parameters of the neural network. The choice of posterior not only allows us to derive an analytical lower bound for the KL divergence between assumed posterior and prior but also gives additional degree of freedom compared to a fully factorized Gaussian.

In order to quantify data uncertainty, we use the standard trick of predicting not only mean but also the variance of survival probability (Kendall & Gal, 2017). Our overall prediction now becomes a sample drawn from this Gaussian, as follows: $(\hat{y}, \hat{\sigma}^2) = f^{\hat{\theta}}(x); y_{out} = \hat{y} + \hat{\sigma}.\epsilon; \epsilon \sim N(0, 1)$. The loss function for the mini batch D^i of our Bayesian variant is given as follows:

$$\hat{\mathcal{L}}(\psi) = -\log p(D^i|\theta^i) + \frac{1}{M}\sum_{i=1}^{N}\left(\frac{\gamma_i}{2}(\mu_i^2 + \sigma_i^2 - \log(\sigma_i^2)) + (1-\gamma_i)\log(\frac{1-\alpha}{1-\gamma_i}) + \gamma_i \log(\frac{\alpha}{\gamma_i})\right) \tag{3}$$

$$\theta^i \sim q_\psi(\theta),$$

where M is the number of mini-batches and N is the total number of parameters. One can see that setting $\alpha = 1$ and $\gamma = 1$ reduces this expression to a fully factorized Gaussian posterior and prior that is used in varational autoencoders (Kingma & Welling, 2013).

We used a simple one-hidden layer Bayesian neural network with spike and slab prior and posterior, and ReLU activation in all but the final layer. Instead of having a fully connected

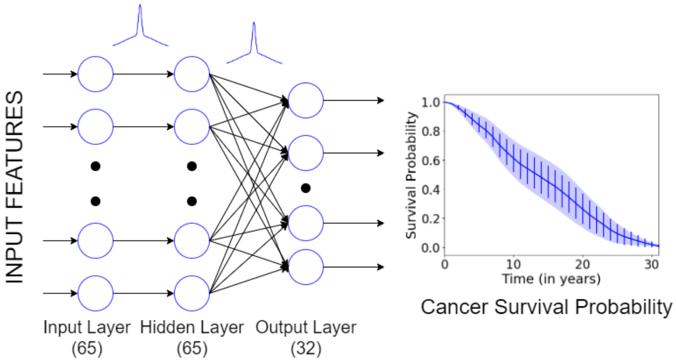

Figure 1: Proposed neural network architecture with weights sampled from the spike and slab posterior give survival probability (solid curve), along with data uncertainty (vertical bars), and model uncertainty (shaded region).

structure from the first layer to the hidden layer, we only assume a one-to-one mapping to simulate variable elimination based on the sparsity inducing prior, as shown in Figure 1.

3 Results

Using a subset of 47 out of the PAM50 gene expressions and clinical variables that were common to both TCGA-BRCA [1] and METABRIC (Curtis et al., 2012) datasets we trained on one dataset and tested on the other to obtain results on model accuracy. We combined both datasets and held out samples at random for experiments on variable importance and uncertainty estimation.

3.1 Survival predictions

C-index and Integrated Brier Score (IBS) are two commonly used metrics for analyzing the accuracy of survival models for censored data, where the former is a generalization of the area under the ROC curve (AUC), and the latter is the average weighted squared distance between the observed and predicted survival. Thus, a higher C-index and lower IBS implies a more accurate model. Table 1 shows our method performs better compared to Cox-PH, MTLR, and a comparable neural-MTLR model with a single hidden layer.

Table 1: Comparison of mean (± std. dev.) C-index and IBS across survival models using one of TCGA-BRCA and METABRIC datasets for training and the other for testing.

Methods	C-index	IBS
Cox-PH	0.65 ± 0.10	0.20 ± 0.07
MTLR	0.68 ± 0.06	0.21 ± 0.06
N-MTLR	0.68 ± 0.02	0.16 ± 0.04
Our Method	0.71 ± 0.05	0.12 ± 0.02

3.2 Ranking prognostic features

We obtained feature importance for each input feature based on the distribution of weights learned by the network from the first layer to the hidden layer. We interpreted the ratio of mean and standard deviation of the weight associated with a feature as its signal to noise ratio. In case of spike and slab posterior, the signal to noise ratio for feature i is given by: $\mid \mu_i \mid / (\sigma_i * \gamma_i)$. We observe in Figure 2 that age at diagnosis, lymph node metastasis, and tumor stage are among the top three prognostically important features.

[1] https://www.cancer.gov/tcga

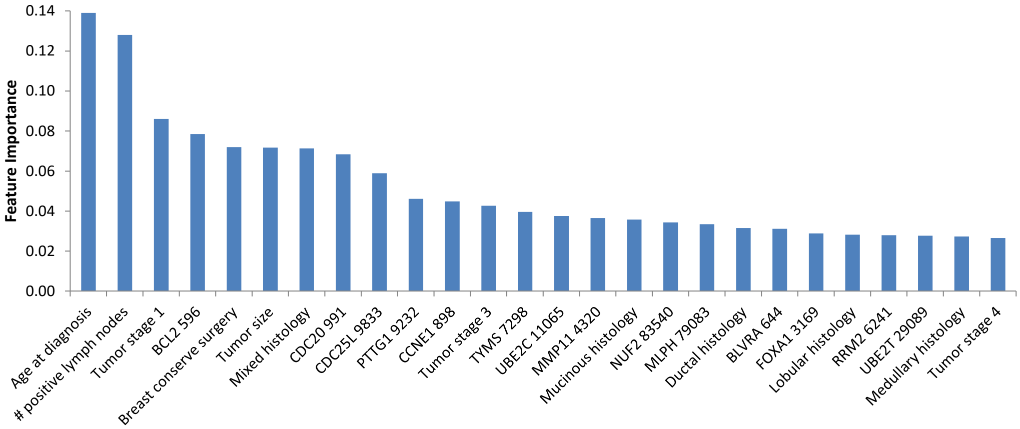

Figure 2: Importance scores for a truncated list of the features (numbers represent gene Entrez ID)

3.3 Low confidence on out of distribution (OOD) test data

In order to demonstrate the use of quantifying uncertainty, we divided the entire data (TCGA + METABRIC) into old (age $>$ 60 years) and young patients (age $<$ 60 years). We trained the model on 80% of the old patients and tested it on the remaining 20% old and all the young patients. We define mean uncertainty score associated with a survival prediction as the mean of the standard deviations in model predictions (for 50 forward passes) across all time points. We observed a 110% higher mean uncertainty score associated with the young patients (OOD) compared to the held-out old patients, as can be seen in Figure 3. Thus, the model was able to identify the pool of young patients as out of the training distribution. We performed similar analysis by training the model on a subset of lower cancer stage patients and saw a 43% higher mean uncertainty score for higher-stage patients (OOD) as compared to the held-out lower stage patients.

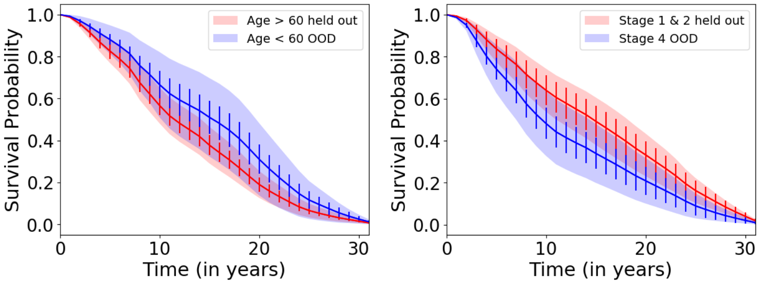

Figure 3: Predicted survival (curve) and model uncertainty (shaded area) for held-out and OOD data.

4 Conclusion

We propose a Bayesian framework for modeling survival prediction that not only gives more accurate predictions but is also interpretable and trustworthy due to its well-calibrated uncertainty estimates. Our model is able to select prognostically important features in the data and detect test samples that are out of the training distribution, making it real-world deployable and capable of producing new biological insights when trained on higher-dimensional data.

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. arXiv preprint arXiv:1505.05424, 2015.

D. R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2):187–220, 1972. ISSN 00359246.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework, 2018.

Karl Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variational free energy and the laplace approximation. Neuroimage, 34(1):220–234, 2007.

Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. Journal of the American Statistical Association, 88(423):881–889, 1993. doi: 10.1080/01621459.1993.10476353.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. Journal of the American Statistical Association, 83(404):1023–1032, 1988. doi: 10.1080/01621459.1988.10478694.

Michalis K. Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), Advances in Neural Information Processing Systems 24, pp. 2339–2347. Curran Associates, Inc., 2011.

Francesco Tonolini, Bjorn Sand Jensen, and Roderick Murray-Smith. Variational sparse coding, 2019. URL https://openreview.net/forum?id=SkeJ6iR9Km.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), Advances in Neural Information Processing Systems 24, pp. 1845–1853. Curran Associates, Inc., 2011.