Uncertainty Estimation in Deep Bayesian Survival Models

Christian Marius Lillelund, Martin Magris and Christian Fischer Pedersen
Department of Electrical and Computer Engineering, Aarhus University
Finlandsgade 22, 8200 Aarhus N, Denmark
Emails: cl@ece.au.dk, magris@ece.au.dk and cfp@ece.au.dk

***Abstract*—Bayesian methods can express uncertainty about their predictions, but have seen little adaptation in survival analysis using neural networks. Proper uncertainty estimation is important in high-risk domains, such as the healthcare or medical field, if machine learning methods are to be adopted for decision-making purposes, however, uncertainty estimation is a known shortcoming of neural networks. In this paper, we introduce the use of Bayesian inference techniques for survival analysis in neural networks that rely on the Cox proportional hazard assumption, for which we discuss a new flexible and effective architecture. We implement three architectures: a fully-deterministic neural network that acts as a baseline, a Bayesian model using variational inference, and one using Monte-Carlo Dropout. Our comprehensive experiments show that on the WHAS500 dataset, Bayesian techniques improve predictive performance over the state-of-the-art neural networks and on the larger SEER and SUPPORT datasets provide comparable performance. In all experiments, training with Monte Carlo Dropout is significantly faster than training with variational inference. Our Bayesian models additionally provide quantification of both aleatoric and epistemic uncertainty, which we exhibit by plotting 95% confidence intervals around the survival function and showing a probability density function of the survival time. Our work motivates further work in leveraging uncertainty for survival analysis using neural networks.**

***Index Terms*—uncertainty estimation, neural networks, survival analysis, variational inference, MC Dropout**

I. Introduction

The Cox's proportional hazards model [1] has long been the standard approach for survival analysis in many healthcare applications, but recent advances in machine learning research have made neural networks (NNs) a powerful tool for survival regression [2]–[4]. Such networks have proven to provide solid performance in terms of ranking and prediction accuracy in survival analysis applications, but traditional maximum-likelihood-based methods notoriously perform poorly when data is sparse [5] and proper uncertainty estimation is a known shortcoming of NNs. In healthcare and biomedical applications, confidence in the prognostic outcome of, e.g., a cancer diagnosis, is essential for medical doctors to make the decision on individual treatment. Bayesian models can express uncertainty about their predictions, specifically aleatoric and epistemic. The former refers to the inherent noise in the data, and the latter to the lack of knowledge of the model [6].

Bayesian Neural Networks (BNNs) are stochastic NNs trained using Bayesian inference, e.g., variational inference (VI) [7] and Monte-Carlo Dropout (MCD) [5]. A BNN was adopted for survival analysis by [8], but their work used pseudo survival probabilities instead of censored ones.

In this work, we propose BNNs as a tool for uncertainty estimation in survival analysis models. To the authors' best knowledge, no prior work has explored the use of BNNs for Cox survival analysis. We consider both VI and MCD as estimation approaches, and, to the end of comparing Bayesian and non-Bayesian approaches, we introduce a feasible network architecture and furthermore adopt five existing reference models. We observe agreement and juxtaposition between deterministic point estimates and the Bayesian ones, indicating the overall validity of our setup. In-depth analyses unveil an advantage of adopting a Bayesian framework, especially in small-sized datasets. The intrinsic probabilistic dimension of BNNs naturally allows for recovering uncertainty estimates, estimating predictive errors and plotting confidence bands. This provides a much more comprehensive description and understanding of the data and model under consideration. Source code is available at: https://github.com/thecml/UE-BNNSurv

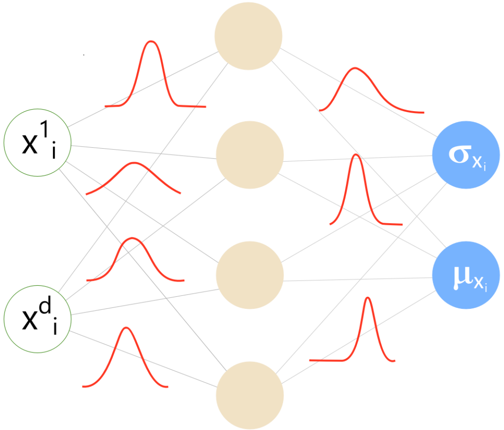

Fig. 1: Illustration of the BNN architecture with a single hidden layer, where weights and biases are treated as random variables. The output nodes provide the mean and standard deviation of a risk score, $\hat{r}_i$, as Gaussian samples, i.e., $\hat{r}_i \sim \mathcal{N}(\mu_{\boldsymbol{x}_i}, \sigma_{\boldsymbol{x}_i})$ for a d-dimensional sample $\boldsymbol{x}_i$.

This work was supported by the PRECISE project under Grant Agreement No. AAL-2021-8-90-CP by the European AAL Association.

II. Fundamentals

A. Elements of survival data analysis and notation

Modelling the probability that an event occurs at time T later than t, i.e., the *survival* probability $S(t) = \Pr(T > t) = 1 - \Pr(t \leq T)$, is a major task in survival analysis. A main ingredient is the so-called hazard function $h(t) = \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t | T > t)/\Delta t$. This corresponds to the death *rate* at an instant after time t, giving survival past that time [9]. The hazard function is related to the survival function through $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density associated with T, $f(t) := \lim_{\Delta t \to 0} \Pr(t < T \leq t + \Delta t)/\Delta t$, i.e., the instantaneous rate of death at time t. In this view, $h(t)$ is the density of T conditional on $T > t$, and the functions $S(t)$, $h(t)$, $f(t)$, all correspond to equivalent ways of describing the distribution of T, formalizing, e.g., the intuition that higher values for $h(t)$ correspond to higher death probabilities.

B. Cox Proportional Hazard Model

For the task of fitting a regression model to survival times, let $\mathcal{D} = \{(y_i, \delta_i, \boldsymbol{x}_i,)\}_{i=1}^N$ be the data, where i denotes the i-th individual. With T_i being the event time and C_i the censoring time, $y_i = \min(T_i, C_i)$, and $\delta_i = 1$ if $T_i \leq C_i$ (zero otherwise). We denote $\boldsymbol{x}_i$ as a vector of d covariates.

Cox's Proportional Hazards (CoxPH) model [1] assumes a conditional individual hazard function of the form $h(t|\boldsymbol{x}_i) = h_0(t)\exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$. The risk score is denoted as $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$. In [1] f is set to a linear function of the covariates, i.e., $f(\boldsymbol{\theta}, \boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\theta}$, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is derived by numerically maximizing the (partial) log-likelihood:

$$\begin{aligned}\log p(\mathcal{D}|\boldsymbol{\theta}) &= \sum_{i:\delta_i=1} \log \frac{h(y_i|\boldsymbol{x}_i)}{\sum_{j:T_j \geq T_i} h(y_i|\boldsymbol{x}_j)}, \\ &= \sum_{i:\delta_i=1} f(\boldsymbol{\theta}, \boldsymbol{x}_i) - \sum_{j:T_j \geq T_i} f(\boldsymbol{\theta}, \boldsymbol{x}_j). \qquad (1)\end{aligned}$$

Different estimators have been derived for the baseline hazard function, e.g., [10], enabling the estimation of the survival function as $\hat{S}(t) = \hat{S}_0(t)^{\exp(\boldsymbol{x}_i\hat{\boldsymbol{\theta}})}$, with $\hat{S}_0(t) = \exp(-\int_0^t \hat{h}_0(t)\mathrm{d}t)$. A purely non-parametric approach, not using covariates' information, is provided by the Kaplan-Meier estimator, used as a reference in Fig. 2. For more details, see, e.g., [9].

C. Variational inference

We denote the data as $\mathcal{D}$, the likelihood as $p(\mathcal{D}|\boldsymbol{\theta})$ and prior distribution on the parameter of interest $\boldsymbol{\theta}$ as $p(\boldsymbol{\theta})$. The target of Bayesian inference is the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})/p(\boldsymbol{\theta})$. The distribution of unobserved (new) data conditioned on $\mathcal{D}$, known as predictive distribution, is obtained from the posterior through marginalization over the parameter space Θ, i.e., $p(\boldsymbol{x}_{\text{new}}|\mathcal{D}) = \int_\Theta p(\boldsymbol{x}_{\text{new}}|\mathcal{D}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}$.

The direct computation of the posterior is challenging as, in general, the term $p(\mathcal{D})$ is intractable. Sampling methods do not scale well in high dimensions and are time-consuming: VI approximates the true and unknown posterior distribution with a distribution $q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})$ chosen within a class of tractable parametric distributions $\mathcal{Q}$. Hereafter $\mathcal{Q}$ is the class of multivariate Gaussians with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\zeta} \equiv \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. VI seeks for the best approximation in $\mathcal{Q}$ to $p(\boldsymbol{\theta}|\mathcal{D})$ by minimizing the Kullback-Leibler (KL) divergence from $q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathcal{D})$, i.e., by recovering the variational parameter $\boldsymbol{\zeta}^\star$ via minimizing:

$$\mathrm{KL}(q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) := \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\zeta}}}\left[\frac{\log p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}|\mathcal{D})}{\log q_{\boldsymbol{\zeta}}(\boldsymbol{\theta})}\right], \qquad (2)$$

commonly attained via the Bayes-by-backprop method [11]. A NN estimated with Bayesian inference is commonly referred to as Bayesian Neural Network (BNN), and our approach provides a BNN-based extension of the non-Bayesian and fully-linear standard Cox model.

D. Monte-Carlo Dropout

Optimizing Eq. (2) requires sampling, estimating stochastic gradients, and involves several caveats and difficulties [6]. Though Monte-Carlo Dropout (MCD) has been historically developed as a computationally efficient method for regularization, it has been shown to have a connection with Bayesian inference. Indeed, NNs with dropout applied at every layer are equivalent to approximate VI in deep Gaussian Processes [12]. Dropout optimizes the KL objective between an approximate distribution and the posterior of a deep Gaussian Process as a mixture of Gaussians [13]. Such a setup is limited compared to VI, yet its simplicity makes it attractive and a widely adopted baseline for comparison with alternative Bayesian approaches. So far, within NN applications in survival analysis, MCD has not been adopted for uncertainty quantification.

III. Experiments and results

A. Network architecture

For the purpose of comparability in the results, consistency in the hyperparameters, and evaluation fairness, we propose a multi-layer perceptron (MLP) backbone network architecture. A MLP is fully-connected feedforward NN; the number of hidden layers and neurons is a hyperparameter tuned based on the data, see, e.g., [6]. Using an MLP is aligned with the literature, as capable of providing solid predictive performance [2]–[4].

The proposed architecture is depicted in Fig. 1. There are three configurations: (i) baseline: a fully-deterministic NN obtained by removing the σ_x neuron and optimizing the objective Eq. (1). (ii) Bayesian model trained with VI, learning the variational approximation by optimizing Eq. (2). (iii) Bayesian model trained with MCD (25% dropout rate).

In relation to Sec. II-C, Eq. (2) adopts the Cox likelihood (1) where $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ corresponds to the MLP neural network and $\boldsymbol{\theta}$ to its parameter (collection of weight and biases), the objective of the Bayesian inference. Whereas the epistemic uncertainty in the model is captured by the Bayesian framework, to capture the aleatoric uncertainty, the network's outputs are sampled from a Gaussian [14], Fig. 1.

B. Datasets and models

For our empirical analyses, we use the openly-available WHAS500 [15], SEER [16] and SUPPORT [17] datasets; they differ in the number of samples and the percentage of censored data (see Tab. I). After imputing missing values by sample mean for real-valued covariates or mode for categorical covariates, applying a z-score data normalization and one-hot encoding categorical covariates, we adopt a 70%-30% train-test split for training and testing.

We implement the Cox model [1], two related traditional survival models CoxNet [18], RSF [19], and three models based on a NN architecture, DSM [3], DeepSurv [2], and the MLP discussed in Sec. III-A. For all datasets and models, we use Bayesian optimization [20] to tune the respective models' hyperparameters over ten iterations using 5-fold cross-validation. This includes number of iterations, batch size and network architecture if applicable, and is done solely on the 70% training split. We use the hyperparameters leading to the highest average concordance-index (Harrell's) on the validation folds.

C. Results

Table I reports the predictive performance of our baseline MLP and its VI and MCD variants, and literature benchmarks. We adopt four evaluation metrics: Harrell's concordance-index (CI_H) [21], Uno's concordance-index (CI_U) [22], the integrated Brier score (IBS) [23] and the negative log-likelihood (NLL). For VI and MCD, such measures are constructed based on predictive means over 100 posterior draws. Model were trained until convergence (no improvement was seen in test CI_H).

Concerning the baseline MLP models and their VI and MCD variants, on the smallest WHAS500 dataset we observe that modelling both aleatoric and epistemic uncertainty by adopting Bayesian methods improves the ranking and accuracy performance in terms of CI_H, CI_U and NLL. In the mid-sized SEER dataset, the effect of the different estimation approaches for the MLP reduces, and all the measures are generally aligned with each other, whereas for the SUPPORT dataset, maximum likelihood can take full advantage of the considerable size of the data and outperforms, consistently although slightly, Bayesian methods.

With respect to the existing models, our work outperforms NN-based solutions on the small dataset (high CI_H, CI_U and low IBS) and we see an advantage in combining the MLP architecture with Bayesian inference for ranking and accuracy in predicting the survival function. In the mid-sized dataset, the performance of all the models is similar, whereas in the large dataset, it seems that methods based on NNs clearly outperform the traditional CoxPH, CoxNet and RSF in terms of CI_H, CI_U and NLL, though behave similarly in terms of IBS.

TABLE I: Performance metrics on the test sets. N: total sample size, C: pct. of censored data, d: number of covariates.

(a) WHAS500 ($N = 500$, $C = 57\%$, $d = 14$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.05s	0.806	0.785	0.152	0.849
CoxNet [18]	0.04s	0.807	0.785	0.149	0.832
RSF [19]	0.07s	0.785	0.754	0.167	
DSM [3]	2.78s	0.785	0.776	0.195	
DeepSurv [2]	0.24s	0.774	0.758	0.185	
This work:					
Baseline (MLP)	10.67s	0.774	0.761	0.154	0.890
+ VI	2m 53s	0.801	0.776	0.142	0.855
+ MCD	6.47s	0.798	0.772	0.160	0.884

(b) SEER ($N = 4024$, $C = 85\%$, $d = 28$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.23s	0.739	0.726	0.105	0.582
CoxNet [18]	0.23s	0.740	0.727	0.105	0.582
RSF [19]	2.05s	0.724	0.714	0.113	
DSM [3]	2.09s	0.738	0.737	0.115	
DeepSurv [2]	0.55s	0.746	0.746	0.113	
This work:					
Baseline (MLP)	24.92s	0.743	0.745	0.107	0.587
+ VI	6m 15s	0.741	0.733	0.109	0.589
+ MCD	26.81s	0.741	0.726	0.110	0.588

(c) SUPPORT ($N = 8873$, $C = 32\%$, $d = 14$).

Model	T_train	CI_H ↑	CI_U ↑	IBS ↓	NLL ↓
CoxPH [1]	0.36s	0.583	0.586	0.222	2.862
CoxNet [18]	0.10s	0.581	0.585	0.224	2.865
RSF [19]	8.81s	0.569	0.571	0.232	
DSM [3]	4.61s	0.603	0.606	0.222	
DeepSurv [2]	0.70s	0.618	0.617	0.221	
This work:					
Baseline (MLP)	58.72s	0.615	0.617	0.210	2.836
+ VI	40m 28s	0.606	0.607	0.310	2.840
+ MCD	60.87s	0.604	0.606	0.216	2.845

Regardless, the use of Bayesian methods does not deteriorate performance metrics in any of our experiments. This demonstrates that aleatoric and epistemic uncertainty can be included in survival analysis using NNs at no additional cost, besides the increase in training time for VI. The leftmost panel in Fig. 2 shows that our methods provide consistent survival function estimates that align with the Kaplan-Meier estimator. VI and MCD additionally capture the aleatoric and epistemic uncertainty in the confidence intervals, which provides a clear predictive advantage and facilitates better adaptation for decision support. In the central panel, observe that grade 1 and 2 tumors have a statistically different impact on survival times approximately after 25 years, a detail that non-probabilistic methods could not capture. This is an ideal application for modelling the underlying uncertainty in, e.g., different treatments. In the rightmost panel, by following [24], we plot the survival time of a randomly selected individual from the test set: the distribution of the exponential random samples with a rate parameter randomly sampled from the MCD predictive distribution is quantitatively different than the one obtained based on the predictive mean, which shows lighter tails, and thus a higher likelihood of early-experiencing the event compared to the non-Bayesian MLP estimation.

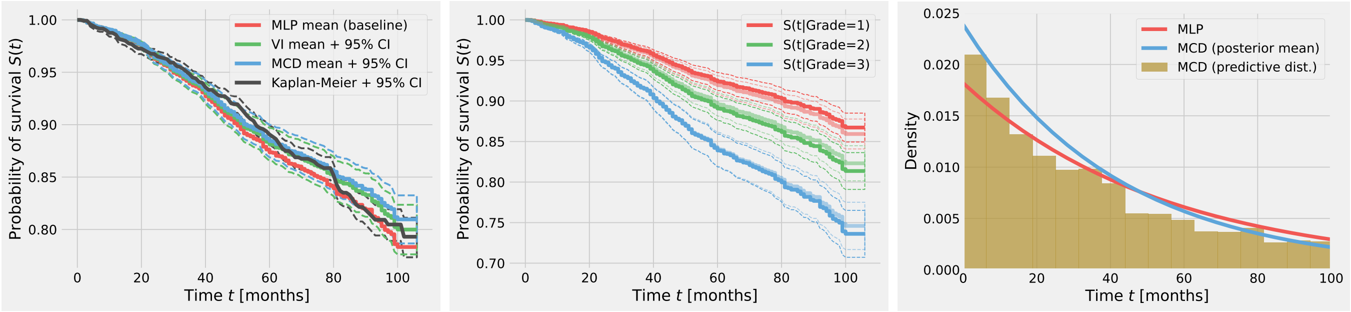

Fig. 2: Model inference on the SEER dataset. Left & center: survival function for different estimation approaches and for various grades of the tumor severity in breast cancer cases (Solid lines: VI; Opaque lines: MCD; Dashed lines: 95% conf. int.). Right: probability density function of the survival time, see Sec. III-C for details.

IV. Conclusion

We adopt variational inference and Monte-Carlo Dropout to provide Bayesian estimation of survival probability and risk in Cox models using neural networks. By proposing a suitable and effective neural network architecture, we perform extensive experiments over six models and three datasets. These show that Bayesian techniques can in situations where data is sparse increase predictive performance of survival models compared to other neural network approaches. We encourage further work within our framework, which we have shown to be effective in providing an immediate quantification of combined aleatoric and epistemic uncertainty. Such information can support risk-aware decision-making in high risk domains, such as the healthcare or medical domain.

References

[1] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[2] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, “Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network,” *BMC Medical Research Methodology*, vol. 18, no. 1, 2018.

[3] C. Nagpal, X. Li, and A. Dubrawski, “Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, 2021.

[4] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, “Deephit: A deep learning approach to survival analysis with competing risks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[5] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016.

[6] M. Magris and A. Iosifidis, “Bayesian learning for neural networks: an algorithmic survey,” *Artificial Intelligence Review*, pp. 1–51, 2023.

[7] A. Graves, “Practical variational inference for neural networks,” in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[8] D. Feng and L. Zhao, “Bdnnsurv: Bayesian deep neural networks for survival analysis using pseudo values,” *Journal of Data Science*, vol. 19, no. 4, pp. 542–554, 2021.

[9] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*, 2nd ed. Spinger, 2021.

[10] N. E. Breslow, “Analysis of survival data under the proportional hazards model,” *International Statistical Review*, pp. 45–57, 1975.

[11] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, 2015, p. 1613–1622.

[12] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 1050–1059.

[13] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] M. Valdenegro-Toro and D. S. Mori, “A deeper look into aleatoric and epistemic uncertainty disentanglement,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 1508–1516.

[15] D. W. Hosmer, S. Lemeshow, and S. May, *Applied survival analysis: regression modeling of time-to-event data*, 2nd ed. Wiley-Interscience, 2008.

[16] L. A. Gloeckler Ries, M. E. Reichman, D. R. Lewis, B. F. Hankey, and B. K. Edwards, “Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) program,” *Oncologist*, vol. 8, no. 6, pp. 541–552, 2003.

[17] W. A. Knaus, “The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults,” *Annals of Internal Medicine*, vol. 122, no. 3, pp. 191–203, 1995.

[18] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for cox’s proportional hazards model via coordinate descent,” *Journal of statistical software*, vol. 39, no. 5, pp. 1–13, 2011.

[19] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[20] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[21] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, “Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors,” *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[22] H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei, “On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data,” *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[23] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, “Assessment and comparison of prognostic classification schemes for survival data,” *Statistics in Medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[24] P. C. Austin, “Generating survival times to simulate cox proportional hazards models with time-varying covariates,” *Statistics in Medicine*, vol. 31, no. 29, pp. 3946–3958, 2012.