Deep Cox Mixtures for Survival Regression

Chirag Nagpal[1,2], **Steve Yadlowsky**[1], **Negar Rostamzadeh**[1] **and Katherine Heller**[1]
CHIRAGN@CS.CMU.EDU
[1] *Brain Team, Google Research*
[2] *Auton Lab, Carnegie Mellon University*

Editor: Editor's name

Abstract

Survival analysis is a challenging variation of regression modeling because of the presence of censoring, where the outcome measurement is only partially known, due to, for example, loss to follow up. Such problems come up frequently in medical applications, making survival analysis a key endeavor in biostatistics and machine learning for healthcare, with Cox regression models being amongst the most commonly employed models. We describe a new approach for survival analysis regression models, based on learning mixtures of Cox regressions to model individual survival distributions. We propose an approximation to the Expectation Maximization algorithm for this model that does hard assignments to mixture groups to make optimization efficient. In each group assignment, we fit the hazard ratios within each group using deep neural networks, and the baseline hazard for each mixture component non-parametrically.

We perform experiments on multiple real world datasets, and look at the mortality rates of patients across ethnicity and gender. We emphasize the importance of calibration in healthcare settings and demonstrate that our approach outperforms classical and modern survival analysis baselines, both in terms of discriminative performance and calibration, with large gains in performance on the minority demographics.

1. Introduction

The importance of survival analysis models in medical applications cannot be overstated. These models support physicians and epidemiologists in clinical decision making based on data-driven evidence about patients' likelihood of survival characteristics based on biological measurements and demographic information about the patients. In this paper, we focus on estimating the patient's risk of an event T of interest, specifically the conditional survival curve, $\mathbb{P}(T > t|X)$. Typically events include death, or the presence or progression of a health condition.

The one frequent challenge with estimating the survival curve is that outcomes are typically censored, meaning that the outcome is unknown for some patients due to lack of follow up or independent competing events. Luckily, censoring is relatively straightforward to deal with in certain commonly used survival analysis models that make the proportional hazards assumption, such as the Cox regression model, or Faraggi-Simon deep neural network model. Unfortunately, in many important cases, the proportional hazards assumption does not hold, leading to poor calibration of patients' estimated survival curve, even if the model can rank patients well.

In fact, many recent deep learning approaches demonstrate significant improvement in ranking patients' survival according to discriminative measures such as the concordance index (C-index). However, the C-index measures pairwise ranking ability and disregards the absolute value of the actual estimated risk score akin to metrics of evaluating binary classification like the Receiver Operation Characteristic.

In this paper we propose, '**Deep Cox Mixtures**' for survival analysis, which generalizes the proportional hazards assumption via a mixture model, by assuming that there are latent groups and within each, the proportional hazards assumption holds. Our approach allows the hazard ratio in each latent group, as well as the latent group membership, to be flexibly modeled by a deep neural network, allowing us to take advantage of the recent improvements in neural network modeling of patient data.

In our experiments, we show that the added flexibility of this mixture of proportional hazards models allows us to improve the calibration of the estimated conditional survival curves, while maintaining excellent discriminative performance; that is, without requiring a performance trade-off. We find that the largest improvements to calibration occur among minority groups, and emphasize the need for evaluating performance on such groups, which can often go unnoticed on dataset-wide performance statistics. Our model is implemented in `tensorflow` and source code of our experiments is open source and publicly available at https://github.com/chiragnagpal/deep_cox_mixtures.

Technical Significance The proposed Deep Cox Mixtures model is not restricted by the strong assumption of proportional hazards. By allowing the model to flexibly choose these latent groups, we can build a more expressive survival analysis model. However inference is challenging owing to the fact that the Cox model involves learning the baseline survival distributions non-parametrically. We develop an approximate Monte Carlo Expectation-Maximization (EM) learning algorithm to estimate the latent groups and parameters of conditional survival curves within each group. To make the learning algorithm tractable, we propose to approximate the Maximization step (M-step) by hard assignment of each patient to a latent group, and approximate the baseline survival curves in the Expectation step (E-step) with spline estimation

Clinical Relevance Survival analysis methods can help healthcare practitioners determine risk, triage and support clinical decision making. However studies show systemic miss-estimation of the prognosis and risk by statistical approaches on some demographics, can lead to wrong and harmful decision making (Vyas et al., 2020). One example is the 2013 ACC/AHA* Pooled Cohort Equations (PCE) to asses cardio-vascular risk (Stone et al., 2014). Yadlowsky et al. demonstrate that 2013 PCEs overestimate the risk for approximately 11.8 million U.S. adults and this overestimation is especially prominent amongst the black population. In this study, we consider improving calibration across minority demographics as a step towards making more equitable models. In particular we reduce the miss-estimation in underrepresented demographics; their calibration and discriminatory performance upon classical and modern survival analysis baselines.

*American College of Cardiology/American Heart Association

2. Related Work

Recent progress in deep learning has also sparked interest in the survival analysis community. Recent thrusts in survival analysis have involved deep learning based Cox models (Katzman et al., 2018) like the original Faraggi-Simon network (Faraggi and Simon, 1995). More recent papers have explored the use of Discrete time models (Lee et al., 2018), recurrent neural architectures (Lee et al., 2019a) as well as fully parametric methods (Nagpal et al., 2020a) for modelling survival outcomes in the presence of censoring. More involved techniques have involved the use of ensembles with black box optimization, auto encoding variational bayes (Chapfuwa et al., 2020; Xiu et al., 2020), as well as adversarial methods (Chapfuwa et al., 2018) to estimate survival outcomes.

Attempts to learn a mixture of Cox models (Nagpal et al., 2019; Rosen and Tanner, 1999) have focused primarily on learning a mixture of log-linear parametric components for the hazard ratio in the partial log-likelihood. These approaches are still subject to the strong assumptions of proportional hazards. Towards the best of our knowledge, our approach is the first attempt at learning a Cox mixture model using the full likelihood and jointly estimating both the parametric relative hazard and the baseline hazard functions. Close lines of work to ours include Chapfuwa et al. (2020); Ranganath et al. (2016), where the authors propose the latent space to be a mixture distribution and sample the outcome event time from a parametric decoder. Our approach differs from this as we do not need to make any strong parametric assumptions on the event outcome times.

There has also been an interest in learning survival models on time-series and temporal data (Lee et al., 2019a). In this paper we restrict our approach to the case with static feature snapshots, although since our approach involves representation learning using neural networks, it can be easily extended to these settings with appropriate choice of recurrent neural networks.

Poor calibration of deep learning methods has been explored recently in machine learning literature (Guo et al., 2017; Nixon et al., 2019). Poor calibration of Deep Learning models in areas like Natural Language Processing (Nguyen and O'Connor, 2015) and Computer Vision has also been demonstrated. Existing lines of research to improve calibration have involved post processing techniques like Platt Scaling, Bayesian and ensemble methods as well as IPM penalties (Kumar et al., 2018) to improve model calibration. Calibration in the specific case of survival models has been an active area of research as well. Lee et al. (2019b) proposed an ensemble of multiple survival analysis models weighted using Black-Box Bayesian optimization for better calibration. This makes for an interesting modelling approach but practical application is challenging due to computational complexity.

Literature in algorithmic fairness has proposed calibration over subgroups as a measure of algorithm fairness (Kleinberg et al., 2016; Chouldechova, 2017; Pleiss et al., 2017). In these works, calibration is typically referred to as 'sufficiency' or 'matching conditional frequencies' (Hardt et al., 2016) and evaluated using reliability diagrams. We stress that as opposed to scenarios where an algorithm is employed to determine the assignment to a service, in healthcare we are typically interested in estimating risk. In as much errors on both sides

(under and over estimation) of the risk are potentially unfair making calibration a well suited metric for fairness evaluation.*

Survival analysis scenarios are also prone to censoring, making estimation of the Expected Calibration Error challenging. Methods involving evaluation for calibration in the presence of censoring have involved simple histogram based binning methods followed by Kaplan-Meier or IPCW estimation of the Survival probability within each bin. More involved recent methods involve non parametric methods like regression splines (Austin et al., 2020) and kernel methods (Yadlowsky et al., 2019). In this tradition, we shine the light on the calibration of models in our empirical evaluations, emphasizing the calibration within minority groups, in particular. We find that without sacrificing discriminative performance, the added flexibility of our mixture model improves calibration, overall and especially in minority groups.

3. The Deep Cox Mixture Model

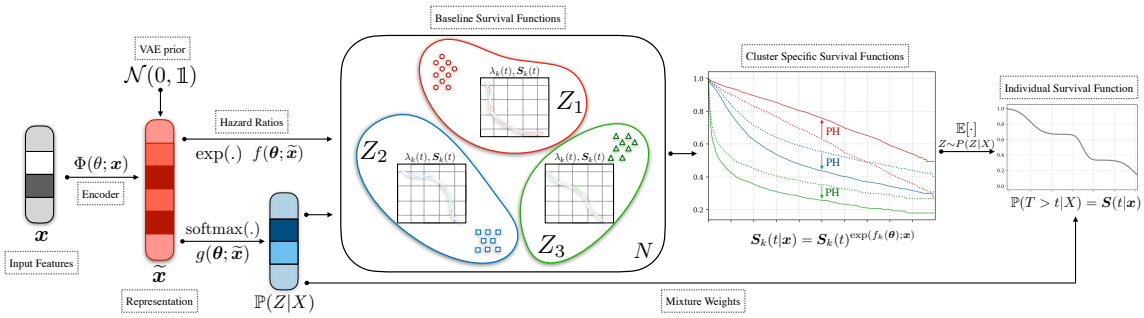


Figure 1: **Deep Cox Mixtures**: Representation of the individual covariates $\boldsymbol{x}$ are generated using an encoding neural network. The output representation $\widetilde{\boldsymbol{x}}$ then interacts with linear functions f and g that determine the proportional hazards within each cluster $Z \in \{1, 2, ...K\}$ and the mixing weights $\mathbb{P}(Z|X)$ respectively. For each cluster, baseline survival rates $\boldsymbol{S}_k(t)$ are estimated non-parametrically. The final individual survival curve $S(t|\boldsymbol{x})$ is an average over the cluster specific individual survival curves weighted by the mixing probabilities $\mathbb{P}(Z|X = \boldsymbol{x})$.

3.1. Notation

We consider a dataset of right censored observations $\mathscr{D} = \{(\boldsymbol{x}_i, \delta_i, u_i)\}_{i=1}^N$ of three tuples, where $\boldsymbol{x}_i$ are the covariates of an individual i, δ_i is an indicator of whether an event occured or not and u_i is either the time of event or censoring as indicated by δ_i.

We consider a maximum likelihood (MLE) based approach to learning $S(t|x) = \mathbb{P}(T > t|X = x)$ from the data. Recall that the survival distribution $S(t|x)$ is isomorphic to the cumulative hazard function $\boldsymbol{\Lambda}(t|x)$, and under continuity, this is equivalent to the hazard function $\boldsymbol{\lambda}(t|x)$. As a result, we will refer them in the parameters of the likelihood

*In healthcare, it is typically ethical to include demographic information like race and gender when estimating outcomes. If there are strong reasons to believe that such information does not cause the outcome, other definitions of algorithmic fairness might be more valid.

interchangeably. Lin (2007) shows that the likelihood of the observed data $\mathscr{D}$ is, up to constant factors,

$$\mathscr{L}(\boldsymbol{\Lambda}) = \prod_{i=1}^{|\mathscr{D}|} \left(\boldsymbol{\lambda}(u_i|\boldsymbol{x}_i)\right)^{\delta_i} \boldsymbol{S}(u_i|\boldsymbol{x}_i). \tag{1}$$

In the following sections, we show how plugging in specific functional forms for $S(t|x)$ allows us to derive survival function estimators.

3.2. MLE for the standard Cox PH model

The key idea behind the Cox model is to assume that the conditional hazard of an individual, is $\boldsymbol{\lambda}(t|x) = \boldsymbol{\lambda}_0(t)\exp\big(f(\boldsymbol{\theta}, x)\big)$, where f is typically a linear function. Under the Cox model, the full likelihood as in equation 1 is

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_0) = \prod_{i=1}^{|\mathscr{D}|} \Big(\boldsymbol{\lambda}_0(u_i)\exp\big(f(\theta, \boldsymbol{x}_i)\big)\Big)^{\delta_i} \boldsymbol{S}_0(u_i)^{\exp\left(f(\boldsymbol{\theta};\boldsymbol{x}_i)\right)} \tag{2}$$

Cox (1972) and the discussion of his paper by Breslow (1972), suggest deriving a maximum likelihood estimate of $\boldsymbol{\theta}$ by maximizing the partial likelihood, $\mathscr{PL}(\boldsymbol{\theta})$ defined below, and using the following estimator of the baseline survival function $\boldsymbol{\Lambda}_0(\cdot)$,

$$\mathscr{PL}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathscr{R}(t_i)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)}, \quad \widehat{\boldsymbol{\Lambda}}_0(t) = \sum_{i:t_i<t} \frac{1}{\sum\limits_{j\in\mathscr{R}(t_i)} \exp\big(f(\widehat{\boldsymbol{\theta}};\boldsymbol{x}_j)\big)}, \tag{3}$$

where $\mathscr{R}(t_i)$ is the 'risk set' – the set of individuals that survived beyond time t_i.

3.3. Proposed Model

In the case of DCM we propose an extension to the Cox model, modeling an individual's survival function using a finite mixture of K Cox models, with the assignment of an individual i to each latent group mediated by a gating function $g(.)$ The full likelihood for this model is

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_k) = \prod_{i=1}^{|\mathscr{D}|} \int_Z \left(\boldsymbol{\lambda}(u_i|\boldsymbol{x}_i)\right)^{\delta_i} \boldsymbol{S}_k(u_i|\boldsymbol{x}_i)\mathbb{P}(Z=k|\boldsymbol{x}_i).$$

$$\text{where, } \boldsymbol{\lambda}(u_i|\boldsymbol{x}_i) = \boldsymbol{\lambda}_k(u_i)\exp\big(f_k(\boldsymbol{\theta}, \boldsymbol{x}_i)\big), \quad \boldsymbol{S}_k(u_i|\boldsymbol{x}_i) = \boldsymbol{S}_k(u_i)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)}$$
$$\text{and, } \mathbb{P}(Z=k|X=\boldsymbol{x}_i) = \text{softmax}\big(g(\boldsymbol{\theta};\boldsymbol{x}_i)\big) \tag{4}$$

Architecture: We allow the model to learn representations for the covariates $\boldsymbol{x}_i$ by passing them through a encoding neural network, $\Phi(.) : \mathbb{R}^d \to \mathbb{R}^h$. This representation then interacts with linear functions f and g defined on $\mathbb{R}^h \to \mathbb{R}^k$; that determine the log hazard ratios and the mixture weights respectively. The set of parameters for the encoder Φ and the linear functions f and g are jointly notated as $\boldsymbol{\theta}$. We experiment with a simple feed forward MLP and a variational auto-encoder for $\Phi(.)$ The parameters of the MLP and the VAE are learnt jointly during learning. For the VAE variant the encoder and the decoder architecture is kept the same. We also experiment with a variant that doesn't use representation learning and thus the functions f and g are linear and restricted to operate on the original features $\boldsymbol{x}$. Figure 1 provides a schematic description of our approach.

3.4. Learning

Notice that under the model in Eq. 4, the corresponding partial likelihood is not independent of $\boldsymbol{\lambda}(.)$, the hazard rate. We hence cannot directly optimize the partial likelihood to perform parameter learning. This inference complexity is outlined in Appendix A.1. Since our model requires inference over the latent assignments Z for learning the Expectation Maximization (Dempster et al., 1977) algorithm is a natural approach to perform inference. The major challenge to applying exact EM lies in the fact that under the our model requires a summation over all possible combinations of latent assignments and which is intractable to compute. We propose an approximate, Monte Carlo EM algorithm (Wei and Tanner, 1990; Song et al., 2016) involving the drawing of posterior samples to learn the parameters, $\boldsymbol{\theta}$ and the baseline survival functions $\{\boldsymbol{S}_k(.)\}_{i=1}^K$.

Algorithm 1: Learning for DCM

Input : Training set, $\mathscr{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)_{i=1}^N\}$;
batches, B;

while *<not converged>* **do**
  for $b \in \{1, 2, ..., B\}$ **do**
    $\mathscr{D}_b \leftarrow$ `sampleMiniBatch` $(\mathscr{D})$
    $\{\gamma_i\}_{i=1}^B \leftarrow$ **E-Step**$(\boldsymbol{\theta}, \{\widehat{\boldsymbol{S}}_k\}_{i=1}^K)$
    $\{\zeta_i\}_{i=1}^B \sim$ Categorical(γ)
    $\boldsymbol{\theta} \leftarrow$ **M-Step**$(\boldsymbol{\theta}, \{\zeta_i, \gamma_i\}_{i=1}^B)$
    for $k \in \{1, 2, ..., K\}$ **do**
      $\widehat{\boldsymbol{S}}_k \leftarrow$ `breslow`$(\boldsymbol{\theta}, \{(t_i, \delta_i)\}_{i=1;\zeta_i=k}^{|\mathscr{D}|})$
      $\widetilde{\boldsymbol{S}}_k \leftarrow$ `splineInterpolate`$(\widehat{\boldsymbol{S}}_k)$
    end
  end
end

Return : learnt parameters, $\boldsymbol{\theta}$;
baseline survival splines $\{\widetilde{\boldsymbol{S}}_k\}_{i=1}^K$

E-Step: Involves estimating the posteriors of Z, $\gamma_i \propto \mathbb{P}(T = t|X, Z)^{\delta_i}\mathbb{P}(T > t|X, Z)^{1-\delta_i}$. The Breslow estimator only gives us the estimates of the survival rates, thus computing the posterior counts, $h_i \propto \mathbb{P}(T = t_i|Z, X)$ for the uncensored instances challenging. We mitigate this by interpolating the Baseline Survival Rate for each latent group, $\boldsymbol{S}_k(.)$ using a polynomial spline. Equation 5 provides the interpolated event probability estimates. (Appendix A.2 describes this in detail.)

$$\begin{aligned}
&\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}_i, Z = k) = \widetilde{\boldsymbol{S}}_k(t)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)} \text{ and,} \\
&\widehat{\mathbb{P}}(T = t|X = \boldsymbol{x}_i, Z = k) = \\
&\qquad - \exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)\frac{\widehat{\mathbb{P}}(T > t|\boldsymbol{x}_i, Z = k))}{\widetilde{\boldsymbol{S}}_k(t)}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)
\end{aligned} \tag{5}$$

Here, $\widetilde{\boldsymbol{S}}_k(t)$ is the baseline survival rate interpolated with a polynomial spline.

M-Step: Once the posterior counts γ_i are obtained, the M-Step involves learning maximizing the corresponding $Q(.)$ function given as

$$\begin{aligned}
Q(\theta) = \sum_{i=1}^{|\mathscr{D}|}\sum_k \gamma_i^k \ln \mathbb{P}(Z|X) + \gamma_i^k \ln \mathbb{P}(t|Z, X); \\
\text{where, } \gamma_i \propto \mathbb{P}(T|X, Z)
\end{aligned} \tag{6}$$

Notice that the γ_i^k are soft counts ($\gamma_i \in [0, 1]$) making parameter inference for the term $\mathbb{P}(T|Z, X)$ intractable. Motivated from Monte-Carlo EM methods We instead sample hard posterior counts $\zeta_i \sim \text{Categorical}(\gamma_i)$.

We replace this with hard posterior counts for the second term, $\ln \mathbb{P}(t|Z, X)$

$$\overline{Q}(\theta) = \sum_{i=1}^{|\mathscr{D}|} \sum_k \gamma_i^k \ln \mathbb{P}(Z|X) + \zeta_i^k \ln \mathbb{P}(t|Z, X);$$
$$\text{where, } \zeta_i \sim \text{Categorical}(\gamma_i) \tag{7}$$

Note that $\mathbb{E}[\overline{Q}(\cdot)] = Q(\cdot)$. Thus, $\overline{Q}(\cdot)$ is an unbiased estimate of the exact $Q(\cdot)$

The first term in $\overline{Q}(\cdot)$ can be optimized using gradient based approaches. The second term can be re-written as a sum over k latent groups variables.

$$\begin{aligned}
\overline{Q}(\theta) &= \sum_{i=1}^{|\mathscr{D}|} \sum_k \gamma_i^k \ln \mathbb{P}(Z|X) + \mathbb{1}\{\zeta_i = k\} \ln \mathbb{P}(t|Z, X) \\
&= \sum_{i=1}^{|\mathscr{D}|} \sum_k \gamma_i^k \ln \mathbb{P}(Z|X) + \sum_{i=1}^{|\mathscr{D}|} \sum_k \mathbb{1}\{\zeta_i = k\} \ln \mathbb{P}(t|Z, X) \\
&= \sum_{i=1}^{|\mathscr{D}|} \sum_k \gamma_i^k \ln \mathbb{P}(Z|X) + \sum_k \sum_{i=1}^{|\mathscr{D}_k|} \ln \mathbb{P}(t|Z, X) \qquad (8) \\
&\qquad \text{(Here, } \mathscr{D}_k \text{ is the set of all } \mathscr{D} \text{ with } \zeta_i = k)
\end{aligned}$$

Now using the fact that the Proportional Hazards assumption holds within each group $\mathscr{D}_k$ we arrive at the form of the $Q(\cdot)$ that we optimize in each minibatch as

$$\widehat{Q}(\boldsymbol{\theta}) = \sum_i^{|\mathscr{D}_b|} \sum_k \gamma_i^k \ln \text{softmax}\big(g(\boldsymbol{\theta}; \boldsymbol{x}_i)\big) + \sum_k \ln \mathscr{PL}(\mathscr{D}_b^k; \boldsymbol{\theta})$$

Here, $\mathscr{D}_b^k$ is the subset of all individuals that have $\zeta_i = k$ within the minibatch b and $\mathscr{PL}(.)$ is the partial likelihood as defined in Equation 3. Thus, the use of hard counts ζ effectively reduces the problem to learning K separate Cox models allowing us to maximize the partial likelihood independently within each $k \in K$.

The parameters of the encoder are also updated during the **M-Step** by adding the loss corresponding to the VAE. Altogether the loss function for optimization is

$$\text{Loss}(\boldsymbol{\theta}; \mathscr{D}_b) = \widehat{Q}(\boldsymbol{\theta}; \mathscr{D}_b) + \alpha \cdot \text{VAE-Loss}(\boldsymbol{\theta}; \mathscr{D}_b) \tag{9}$$

Here, the VAE-Loss is the Evidence Lower Bound for the VAE with representations drawn from a zero mean and identity covariance gaussian prior as in Kingma and Welling (2013).

Algorithm 1 describes the learning procedure for DCM. We sample minibatches $\mathscr{D}_b$ from the data $\mathscr{D}$ and compute the soft and hard posterior counts, $\{\gamma_i, \zeta_i\}_{i \in \mathscr{D}_b}$ for each batch. This is followed by the **M-Step** involving a gradient update the parameter set $\boldsymbol{\theta}$. Finally, we update the Baseline Survival Splines, $\widetilde{\boldsymbol{S}}_k$ computed using the Breslow's estimator (Eq. 3) for each cluster. Note that the Breslow's estimator is computed over the full batch, $\mathscr{D}$. This is computed analytically, does not involve gradient computation and so is not expensive.

3.5. Inference

Following Equation 4, at test time the estimated risk of an individual at time t is given as

$$\begin{aligned}\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}_i) &= \mathbb{E}_{Z\sim\widehat{\mathbb{P}}(Z|X)}[\widehat{\mathbb{P}}(T|X = \boldsymbol{x}_i, Z)] \\ &= \sum_k \widetilde{\boldsymbol{S}}_k(t)^{\exp\left(f(\boldsymbol{\theta};\boldsymbol{x}_i)\right)} \times \text{softmax}_k\big(g(\boldsymbol{\theta}; x_i)\big) \end{aligned} \tag{10}$$

4. Experiments

Table 1: Summary statistics for the datasets used in the experiments.

Dataset	N	d	Censoring (%)	Minority Class (%)	Event Quantiles		
					$t = 25$th	$t = 50$th	$t = 75$th
SUPPORT	9,105	44	31.89%	Non-White (21.02%)	14	58	252
FLCHAIN	6,524	8	69.93%	Female (44.94%)	903.25	2085	3246
SEER	55,993	168	72.82%	Non-White (23.77%)	25	55	108

In this section we describe the datasets, the survival analysis tasks and baselines we compare DCM against. We also describe the corresponding metrics we employ for evaluation.

4.1. Datasets

We experiment with the following real world, publicly available survival analysis datasets:

FLCHAIN (Assay of Serum Free Light Chain): This is a public dataset introduced by Dispenzieri et al. (2012) aiming to study the relationship between serum free light chain and mortality. It includes covariates like age, gender, serum creatinine and presence of monoclonal gammapothy. We removed all the individuals with missing covariates and experiment with the remaining subset of 6,524 individuals. Out of this subset 45% of the participants were coded as female and are considered as 'minority' in our experiments.

SUPPORT (Study to understand prognoses and preferences for outcomes and risks of treatments (Connors et al., 1995): Dataset from study instituted to understand patient survival for 9,105 terminally ill patients on life support. The median survival time for the patients in the study was 58 days. Out of the $9,105$ patients a majority 79% were coded as 'White', while the rest were coded as 'Black', 'Hispanic' and 'Asian'.

SEER (Surveillance, Epidemiology and End Results Study)[†] : This dataset from National Cancer Institute (2019) consists of survival characteristics of oncology patients taken from cancer registries covering about one-third of the US Population. For our study we consider a cohort of patients over a 15 year period from 1992-2007 diagnosed with breast cancer with a

[†] https://seer.cancer.gov/

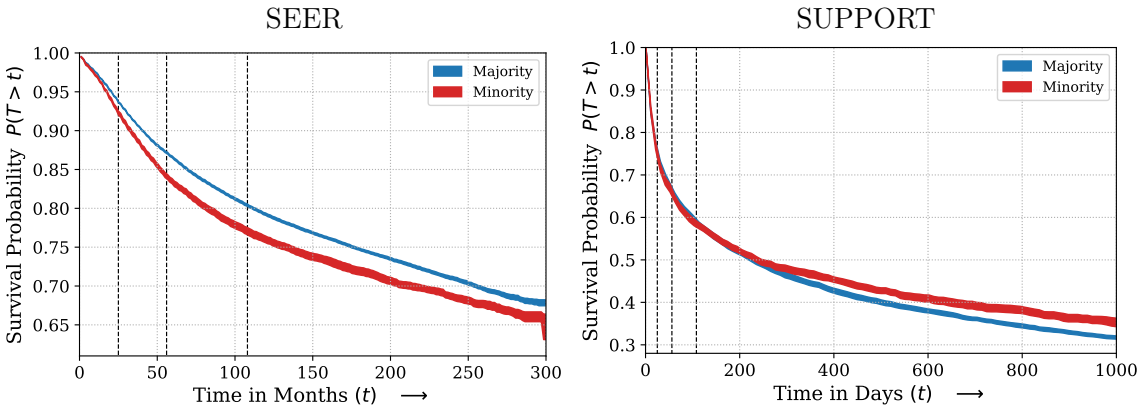


Figure 2: Base survival rates for the majority (White) vs. the other demographics in the SEER dataset estimated with a Kaplan-Meier estimator. Notice that the baseline survival rates differ across groups. Dashed lines respresent the $25^{\text{th}}, 50^{\text{th}}$ and 75^{th} event quantiles.

median survival time of 55 months. A majority (76%) of the patients were coded as 'White' and the rest were other minorities consisting of 'Blacks', 'American Indians', 'Asians', etc.[‡]

Our choice of datasets encompass varying ranges of dimensionality of covariates, levels of censoring and size vis-a-vis the minority demographics. Table 1 describes some summary statistics of the considered datasets. Figure 2 compares the baseline survival rates for the majority and minorities in the SEER and SUPPORT dataset. Notice that base survival rates across demographics can vary considerably over time.

4.2. Baselines

We compare the proposed DCM against the following baselines.

Accelerated Failure Time (AFT): This is an extension of generalized linear models to the survival setting with censored data. The target variable is assumed to follow a Weibull distribution and the shape and scale parameters are modelled as linear functions of the covariates. Parameter learning is performed using Maximum Likelihood Estimation.

Deep Survival Machines (DSM) (Nagpal et al., 2020a): This is another fully parametric approach and improves on the Accelerated Failure Time model by modelling the event time distribution as a fixed size mixture over Weibull or Log-Normal distributions. The individual mixture distributions are themselves parametrized with neural networks allowing to learn complex non-linear representations of the data.

Deep Hit (DHT) (Lee et al., 2018): A discrete time model, DeepHit is a popular Neural Network approach that involves discretizing the event outcome space and treating the survival analysis problem as a multiclass classification problem over the discrete intervals.

[‡]SEER has a very intricate coding pattern vis-a-vis race. Refer to https://seer.cancer.gov/tools/codingmanuals/race_code_pages.pdf for details.

Cox Proportional Hazards (CPH): CPH assumes that individuals across the population have constant proportional hazards overtime.

Faraggi-Simon Net (FSN)/DeepSurv (Faraggi and Simon, 1995; Katzman et al., 2018): An extension to the CPH model, FSN involves modelling the proportional hazard ratios over the individuals with Deep Neural Networks allowing the ability to learn non linear hazard ratios.

Random Survival Forest (RSF) (Ishwaran et al., 2008): RSF is an extension of Random Forests to the survival settings where risk scores are computed by creating Nelson-Aalen estimators in the splits induced by the Random Forest.

Note: In practice we observe that performance of the **Random Survival Forest** model, especially in terms of calibration is strongly influenced by the choice for the hyper-parameters, `mtry` (the number of features considered at each split) and `min_node_size` (the minimum number of data samples to continue growing a tree). We thus advise carefully tuning these hyper-parameters while benchmarking **RSF**.

The full set of hyper parameter we perform grid search on is deferred to Appendix C.

4.3. Evaluation Metrics

We compare the performance of DCM against baselines in terms of both discriminative performance and calibration using the following metrics:

Area under ROC Curve (AUC): Involves treating the survival analysis problem as binary classification at different quantiles of event times and computing the corresponding area under the ROC curve.

Time Dependent Concordance Index (C^{td}): Concordance Index estimates ranking ability by exhaustively comparing relative risks across all pairs of individuals in the test set. We employ the 'Time Dependent' variant of Concordance Index that truncates the pairwise comparisons to the events occurring within a fixed time horizon.

$$C^{td}(t) = \mathbb{P}\big(\hat{F}(t|\mathbf{x}_i) > \hat{F}(t|\mathbf{x}_j)|\delta_i = 1, T_i < T_j, T_i \leq t\big)$$

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $R(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\text{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|R(t)) - R(t)\big|\big]$$

by partitioning the risk scores R into q quantiles $\{[r_j, r_{j+1})\}_{j=1}^{q}$.

Brier Score (BS): The Brier Score involves computing the Mean Squared Error around the binary forecast of survival at a certain event quantile of interest. Brier Score is a proper

scoring rule and can be decomposed into components that measure both discriminative performance and calibration.

$$\text{BS}(t) = \mathbb{E}_{\mathscr{D}}\left[\left(\mathbb{1}\{T > t\} - \widehat{\mathbb{P}}(T > t|X)\right)^2\right]$$

Each of the metrics described above are adjusted for censoring by using standard Thompson-Horvitz style Inverse Propensity of Censoring Weights (IPCW) estimates learnt with a Kaplan-Meier estimator over the censoring times. Details are in Appendix B.

4.4. Experimental Protocol

For the proposed model, DCM and the baselines we perform 5-fold cross validation. The predictions of each fold at the 25th, 50th and 75th quantiles of event times are collapsed together and bootstrapped in order to generate standard errors. For the proposed model and the baselines we report the mean of the evaluation metric and the bootstrapped[§] standard errors for the model that has the lowest Brier Score amongst all the competing set of hyper parameter choices. For DCM, the set of hyperparameter choices include the number of hidden layers for Φ tuned from $\{1, 2\}$, units in each hidden layer selected from $\{50, 100\}$, the number of mixture components K which are tuned between $\{3, 4, 6\}$ and the discounting factor for the VAE-Loss, α tuned from $\{0, 1\}$. Optimization is performed using the Adam optimizer (Kingma and Ba, 2014) in `tensorflow` with learning rates fixed 1×10^{-3} and mini batch size of `128`. The Baseline Survival Splines are fixed to be of degree `3` and fit using the `scipy` python package.

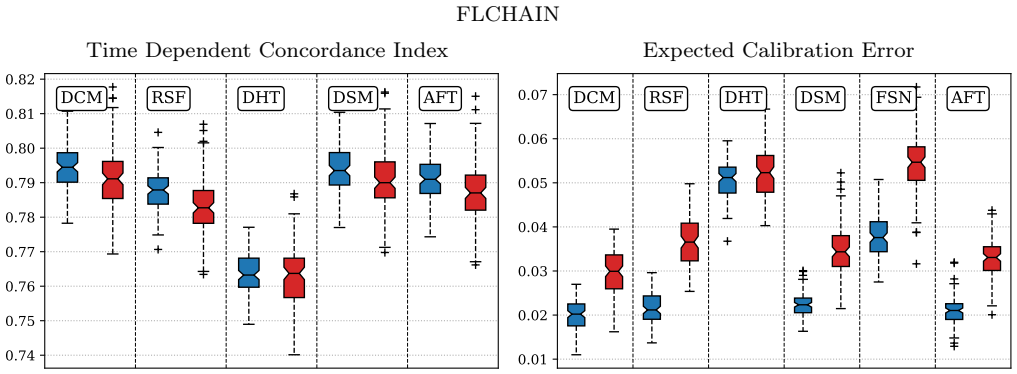
5. Results

In this section we describe the results of our various experiments with DCM and the competing baselines. We present the discriminative performance and calibration for DCM against the baselines on the three datasets for the entire population as well as the minority demographic on the 75^{th} quantile of event times in Figures 3, 4 and 5 and the corresponding tables. (For tabulated results including AuROC and Brier Scores, refer to D.1.)

FLCHAIN: DCM beat all the other baselines in terms of discriminative performance on the entire population as well as on the minority, 'Female' subgroup. In terms of calibration DCM was also consistently better than all the other baselines as evidenced from low ECE scores. Interestingly, both the FSN and the linear Cox model did poorly in terms of concordance and calibration while DCM had good performance suggesting it is not sensistive to proportional hazards (PH).

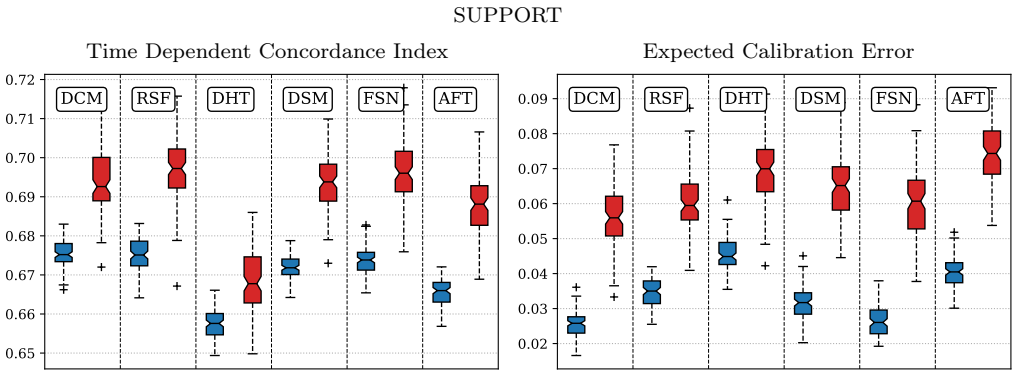
SUPPORT: For the SUPPORT dataset, RSF had the best discriminative performance at a population level, and DCM came a close and beat the other deep learning baselines. Interestingly we found that the proposed DCM had the best discriminative performance on the minority demographic beating all other baselines including RSF.

[§] 100 times



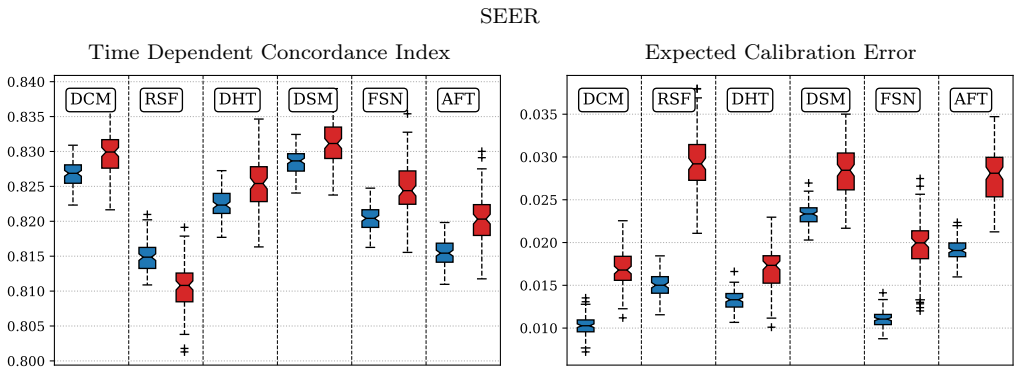
Model	C^{td}		ECE	
	Population	Minority	Population	Minority
CPH	0.6621 ± 0.0087	0.6737 ± 0.0124	0.0992 ± 0.0044	0.0878 ± 0.0071
AFT	0.7911 ± 0.0060	0.7875 ± 0.0087	0.0212 ± 0.0034	0.0329 ± 0.0046
RSF	0.7880 ± 0.0059	0.7830 ± 0.0089	0.0215 ± 0.0037	0.0368 ± 0.0053
FSN	0.6608 ± 0.0081	0.6212 ± 0.0131	0.0381 ± 0.0046	0.0545 ± 0.0068
DHT	0.7636 ± 0.0059	0.7631 ± 0.0092	0.0505 ± 0.0041	0.0525 ± 0.0056
DSM	0.7937 ± 0.0061	0.7909 ± 0.0087	0.0223 ± 0.0029	0.0347 ± 0.0056
DCM	0.7943 ± 0.0103	0.7911 ± 0.0091	0.0200 ± 0.0034	0.0294 ± 0.0049

Figure 3: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75$^{\text{th}}$ event quantile for FLCHAIN.



Model	C^{td}		ECE	
	Population	Minority	Population	Minority
CPH	0.6686 ± 0.0034	0.6905 ± 0.0078	0.0310 ± 0.0041	0.0685 ± 0.0079
AFT	0.6657 ± 0.0034	0.6883 ± 0.0078	0.0402 ± 0.0046	0.0741 ± 0.0085
RSF	0.6751 ± 0.0040	0.6974 ± 0.0084	0.0348 ± 0.0041	0.0603 ± 0.0080
FSN	0.6736 ± 0.0037	0.6961 ± 0.0074	0.0262 ± 0.0040	0.0601 ± 0.0097
DHT	0.6575 ± 0.0038	0.6680 ± 0.0088	0.0457 ± 0.0044	0.0696 ± 0.0089
DSM	0.6718 ± 0.0033	0.6939 ± 0.0079	0.0315 ± 0.0047	0.0650 ± 0.0087
DCM	0.6753 ± 0.0036	0.6939 ± 0.0079	0.0256 ± 0.0037	0.0561 ± 0.0085

Figure 4: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75$^{\text{th}}$ event quantile for SUPPORT.



Model	C^{td}		ECE	
	Population	Minority	Population	Minority
CPH	0.8082 ± 0.0020	0.8121 ± 0.0037	0.0718 ± 0.0015	0.0764 ± 0.0028
AFT	0.8155 ± 0.0020	0.8204 ± 0.0035	0.0192 ± 0.0011	0.0278 ± 0.0029
RSF	0.8153 ± 0.0021	0.8105 ± 0.0035	0.0147 ± 0.0013	0.0270 ± 0.0029
FSN	0.8204 ± 0.0019	0.8248 ± 0.0036	0.0119 ± 0.0011	0.0196 ± 0.0029
DHT	0.8224 ± 0.0020	0.8255 ± 0.0037	0.0133 ± 0.0012	0.0170 ± 0.0024
DSM	0.8281 ± 0.0019	0.8243 ± 0.0036	0.0259 ± 0.0014	0.0311 ± 0.0031
DCM	0.8270 ± 0.0019	0.8296 ± 0.0034	0.0103 ± 0.0011	0.0169 ± 0.0024

Figure 5: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at the 75^{th} event quantile for SEER.

While RSF was strong in terms of discriminative performance, it was however poorly calibrated in comparison to other baselines. DCM had the lowest ECE at each quantile amongst all baselines, at both the population level as well as on the minority demographic. The performance of FSN was close to DCM in terms of calibration but did poorly in terms of discrimination, further lending evidence to the fact that DCM is not restricted by PH.

SEER: In terms of calibration DCM beat all the other baselines at all quantiles of interest for the entire population as well as for the minority group. DCM also consistently had good discriminative power. (DSM did slightly better in terms of discrimination at a population level, but this was not significant). We found that DHT was a strong competitor, which is understandable since it is particularly well suited for discrete time datasets like SEER. Note that in the case of SEER we also report results stratified by the four largest minority demographics in the subset of the dataset we work with in Figure 7. DCM has better discriminative performance across groups especially at longer horizons of event times. In terms of calibration, DCM comes close to or outperforms the semi-parametric approaches like FSN/DeepSurv.

In order to assess the influence of the protected attribute to determine the outcome we conduct additional studies of DCM on the SEER dataset involving removal of the protected attribute (unawareness). We find that unawareness results in overall poorer discriminative performance and calibration. Although unaware models were better calibrated at shorter time horizons, suggesting non-monotonic interaction of group attribute vis-a-vis calibration. These results are deferred to Appendix D.2.

5.1. Learnt Latent Groups

For the SUPPORT dataset, we run DCM with the default set of hyperparameters with $k = 3$ latent groups. We compare the subgroup level survival curves for the learnt subgroups using DCM by plotting the mean survival rates within each group estimated with DCM as well as a Kaplan Meier Estimator.

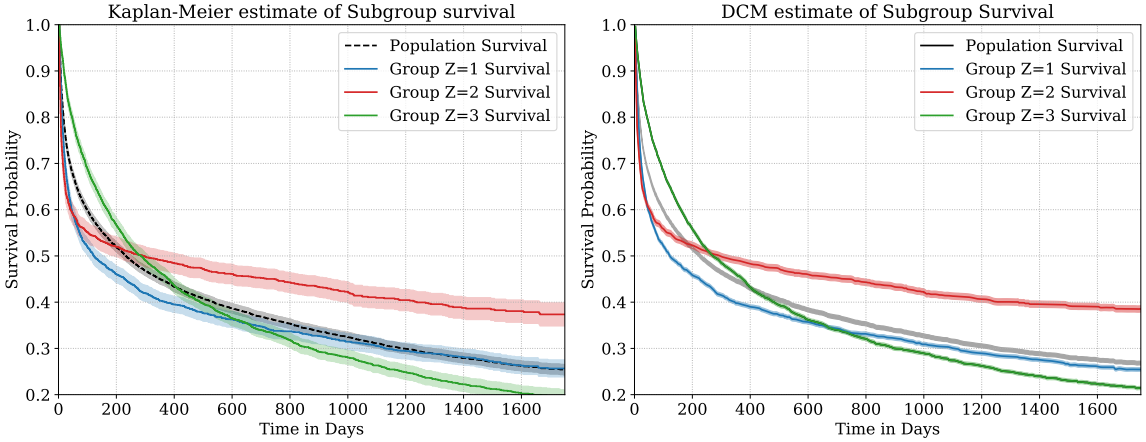


Figure 6: Group specific baseline survival rates for the estimated subgroups using DCM with $k = 3$ for a heldout fold of the SUPPORT dataset. The first plot is the group specific Kaplan-Meier plot and the second plot is the survival estimated with DCM.

Figure 6 present the estimate survival rates of the discovered subgroups using a Kaplan-Meier curve and DCM respectively. Consider the subgroup level survival curves for groups $Z = 2$ and $Z = 3$ that intersect. Intersecting survival curves indicate non Proportional Hazards which DCM is able to capture.

6. Ethical considerations in survival prediction*

Survival prediction and disease prognosis models can help on advanced care decision making and giving recommendations to the patients and their caregivers. Following the model card recommendations (Mitchell et al., 2019), we would like to discuss the intended uses of this technology, and that, using survival prediction techniques, healthcare providers can get more information on the treatment options, and in particular, preventive interventions and their potential outcomes given the predictive models and circumstantial characteristics.

Our models are developed as a proof of concept and the existing datasets including the datasets used in these studies still lack some confounding characteristics that may be causally related to the final outcome (Gaille et al., 2020). For example, studies show black women with breast cancer in the US have a higher mortality rate than white women (Yedjou et al., 2019). However, there is an ongoing discussion on if the mortality rate is related to hereditary (Yedjou et al., 2019) or other factors such as late diagnosis due to the historical discrimination or a combination of all these factors (George et al., 2015). In addition, it is important to note that there should be extensive considerations in using survival analysis systems.

First, these systems should not be used to make decisions in a fully automated manner. There are many reasons for this including well documented biases embedded in historic data. These systems are also not intended to be used in cases of scarce resources, such as ventilators in Covid-19 cases, to rank patients by likelihood of survival[¶] (Beil et al., 2019). Many people in the disability justice space have discussed how disabled patients are often discriminated against based on quality of life estimates . It is important to note that estimated survival rates, should not be the only means in decision making, and quality of life as a factor should be discussed with patients and their families.

In addition, it is very important to know where and how these predictions are used and who will have access to these analyses. For example, the outcome is not well-suited for usage in insurance policies premium (Chiang, 1984; Czado and Rudolph, 2002), patient ranking and immigration status recommendations and not as an automatic means of decision making. In this work we address the issues of result disparity for underrepresented groups, however, we want to state that not all factors are presented in this study that may potentially contribute to the system predictions. Finally, we stress that the causal association of protected attributes like 'Race' and 'Gender' with outcomes is largely problem dependent and an open research problem . Given that 'Race', 'Gender' are not clearly defined attributes even in medical contexts, we urge readers to exercise caution and best judgement when choosing to include these attributes to model outcomes.

7. Conclusion

We proposed 'Deep Cox Mixtures' to model censored Time-to-Event data. Our approach involves estimating hazard ratios within latent clusters followed by non-parametric estimation of the baseline survival rates but is not limited by the strong assumptions of constant proportional hazards. We experiment with several real-world health datasets and demonstrate superiority of our approach both in terms of discriminative performance and calibration with an emphasis on improvements especially on the minority demographics.

In the future, we aim to apply Deep Cox Mixtures to real world health problems to phenotype and recover patients stratified by their relative risk profiles with the overall goal of actionable decision support for clinicians (Wang and Rudin, 2017; Ustun and Rudin, 2019; Chapfuwa et al., 2020). Future extensions can also involve explicitly modelling the effect of an intervention or treatment at an individual (Chapfuwa et al., 2021) or subgroup level (Nagpal et al., 2020b) for retrospective analysis of studies with censored outcomes.

References

Peter C Austin, Frank E Harrell Jr, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ici) for survival models. *Statistics in Medicine*, 2020.

Michael Beil, Ingo Proft, Daniel van Heerden, Sigal Sviri, and Peter Vernon van Heerden. Ethical considerations about artificial intelligence for prognostication in intensive care. *Intensive Care Medicine Experimental*, 7(1):70, 2019.

[¶] https://www.healthaffairs.org/do/10.1377/hblog20200911.401376/

Norman E Breslow. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34:216–217, 1972.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. *arXiv preprint arXiv:1804.03184*, 2018.

Paidamoyo Chapfuwa, Chunyuan Li, Nikhil Mehta, Lawrence Carin, and Ricardo Henao. Survival cluster analysis. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 60–68, 2020.

Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 133–145, 2021.

Chin Long Chiang. *The life table and its applications*. Krieger Malabar, FL, 1984.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Alfred F Connors, Neal V Dawson, Norman A Desbiens, William J Fulkerson, Lee Goldman, William A Knaus, Joanne Lynn, Robert K Oye, Marilyn Bergner, Anne Damiano, et al. A controlled trial to improve care for seriously iii hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (support). *Jama*, 274(20):1591–1598, 1995.

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

Claudia Czado and Florian Rudolph. Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics*, 31(3):395–413, 2002.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pages 517–523. Elsevier, 2012.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Marie Gaille, Marco Araneda, Clément Dubost, Clémence Guillermain, Sarah Kaakai, Elise Ricadat, Nicolas Todd, and Michael Rera. Ethical and social implications of approaching death prediction in humans-when the biology of ageing meets existential issues. *BMC Medical Ethics*, 21(1):1–13, 2020.

Prethibha George, Sheenu Chandwani, Molly Gabel, Christine B Ambrosone, George Rhoads, Elisa V Bandera, and Kitaw Demissie. Diagnosis and surgical delays in african american and white women with early-stage breast cancer. *Journal of Women's Health*, 24(3): 209–217, 2015.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814, 2018.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019a.

Changhee Lee, William Zame, Ahmed Alaa, and Mihaela Schaar. Temporal quilting for survival analysis. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 596–605, 2019b.

DY Lin. On the breslow estimator. *Lifetime data analysis*, 13(4):471–480, 2007.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Chirag Nagpal, Rohan Sangave, Amit Chahar, Parth Shah, Artur Dubrawski, and Bhiksha Raj. Nonlinear semi-parametric models for survival analysis. *arXiv preprint arXiv:1905.05865*, 2019.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. 2020a.

Chirag Nagpal, Dennis Wei, Bhanukiran Vinzamuri, Monica Shekhar, Sara E Berger, Subhro Das, and Kush R Varshney. Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 19–29, 2020b.

Surveillance Research Program National Cancer Institute, DCCPS. Surveillance, epidemiology, and end results (seer) program research data (1975-2016), 2019. URL www.seer.cancer.gov.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, pages 38–41, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

Ori Rosen and Martin Tanner. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9):1119–1131, 1999.

Zhao Song, Ricardo Henao, David Carlson, and Lawrence Carin. Learning sigmoid belief networks via monte carlo expectation maximization. In *Artificial Intelligence and Statistics*, pages 1347–1355, 2016.

Neil J Stone, Jennifer G Robinson, Alice H Lichtenstein, C Noel Bairey Merz, Conrad B Blum, Robert H Eckel, Anne C Goldberg, David Gordon, Daniel Levy, Donald M Lloyd-Jones, et al. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2889–2934, 2014.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *J. Mach. Learn. Res.*, 20: 150–1, 2019.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

Tong Wang and Cynthia Rudin. Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*, 2017.

Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Zidi Xiu, Chenyang Tao, and Ricardo Henao. Variational learning of individual survival distributions. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 10–18, 2020.

Steve Yadlowsky, Rodney A Hayward, Jeremy B Sussman, Robyn L McClelland, Yuan-I Min, and Sanjay Basu. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of internal medicine*, 169(1):20–29, 2018.

Steve Yadlowsky, Sanjay Basu, and Lu Tian. A calibration metric for risk scores with survival data. In *Machine Learning for Healthcare Conference*, pages 424–450, 2019.

Clement G Yedjou, Jennifer N Sims, Lucio Miele, Felicite Noubissi, Leroy Lowe, Duber D Fonseca, Richard A Alo, Marinelle Payton, and Paul B Tchounwou. Health and racial disparity in breast cancer. In *Breast Cancer Metastasis and Drug Resistance*, pages 31–49. Springer, 2019.

Supplementary Materials

Appendix A. Additional details on DCM implementation

A.1. Non Applicability of the Partial Likelihood for the Proposed Model

In this section, we demonstrate that we cannot directly maximize the partial likelihood to learn our model. In the case of the Cox model, the hazard rate for an individual with covariates $\boldsymbol{x}_i$ at time t, $\boldsymbol{\lambda}(t|\boldsymbol{x}_i)$ is given as

$$\boldsymbol{\lambda}(t|\boldsymbol{x}_i) = \boldsymbol{\lambda}_0(t)\exp(f(\beta, \boldsymbol{x}_i)).$$

Here, $\boldsymbol{\lambda}_0(t)$ is the baseline hazard. Now the partial likelihood $\mathscr{PL}(\boldsymbol{\theta})$ is defined as

$$\mathscr{PL}(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\boldsymbol{\lambda}(t|\boldsymbol{x}_i)}{\sum\limits_{j\in\mathscr{R}(t_i)} \boldsymbol{\lambda}(t|\boldsymbol{x}_j)} = \prod_{i:\delta_i=1} \frac{\cancel{\boldsymbol{\lambda}_0(t)}\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathscr{R}(t_i)} \cancel{\boldsymbol{\lambda}_0(t)}\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)} \tag{11}$$

$$= \prod_{i:\delta_i=1} \frac{\exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_i)\big)}{\sum\limits_{j\in\mathscr{R}(t_i)} \exp\big(f(\boldsymbol{\theta};\boldsymbol{x}_j)\big)}. \tag{12}$$

Under our model, the hazard rate for an individual with covariates $\boldsymbol{x}_i$ at time t, $\boldsymbol{\lambda}(t|\boldsymbol{x}_i)$ is given as

$$\boldsymbol{\lambda}(\cdot|\boldsymbol{x}_i) = \frac{\mathbb{P}(t|\boldsymbol{x}_i)}{\boldsymbol{S}(t|\boldsymbol{x}_i)} = \frac{\sum\limits_k \mathbb{P}(t|\boldsymbol{x}_i, Z=k)\mathbb{P}(Z=k|\boldsymbol{x}_i)}{\sum\limits_k \boldsymbol{S}(t|\boldsymbol{x}_i, Z=k)\mathbb{P}(Z=k|\boldsymbol{x}_i)}$$

Clearly, we do not have the proportional hazards form for DCM and so cannot directly optimize the Partial Likelihood independent of the baseline hazard rate.

A.2. Spline Estimates

We want to extract the probabilities estimates $\mathbb{P}(T|Z,X)$ in order to compute the posterior $\mathbb{P}(Z|T,X) \propto \mathbb{P}(T|Z,X)$ for the uncensored observations. We only have access to the estimated survival function from the Breslow's estimate, $\widehat{\boldsymbol{S}}(T > t|X = \boldsymbol{x}_i)$.

$$\begin{aligned}
\mathbb{P}(T > t|X=\boldsymbol{x}_i, Z=k) &= 1 - \mathbb{P}(T \le t|X=\boldsymbol{x}_i, Z=k) \\
&= 1 - \text{cdf}(T \le t|X=\boldsymbol{x}_i, Z=k) \\
\text{Now, } \text{cdf}(T \le t|X=\boldsymbol{x}_i, Z=k) &= 1 - \mathbb{P}(T > t|X=\boldsymbol{x}_i, Z=k) \\
\frac{\partial}{\partial t}\text{cdf}(T \le t|X=\boldsymbol{x}_i, Z=k) &= \frac{\partial}{\partial t}\Big(1 - \mathbb{P}(T > t|X=\boldsymbol{x}_i, Z=k)\Big) \qquad \text{[taking derivative wrt. } t\text{]} \\
\implies \text{pdf}(T=t|X=\boldsymbol{x}_i, Z=k) &= -\frac{\partial}{\partial t}\mathbb{P}(T > t|X=\boldsymbol{x}_i, Z=k) \\
&= -\frac{\partial}{\partial t}\boldsymbol{S}_k(t)^{\exp(f_k(\boldsymbol{\theta};\boldsymbol{x}_i))}
\end{aligned}$$

Here pdf$(\cdot)$ and cdf$(\cdot)$ are the probability density and the cumulative density functions respectively. Now replacing the baseline survival function $\boldsymbol{S}_k(.)$ with the interpolated spline estimate, $\widetilde{\boldsymbol{S}}_k(.)$ we get the spline estimate of $\mathbb{P}(T = t|Z, X)$ as

$$\begin{aligned}\widehat{\mathbb{P}}(T = t|Z, X) &= -\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)^{\exp(f_k(\boldsymbol{\theta};\boldsymbol{x}_i))} \\ &= -\exp\big(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\big)\widetilde{\boldsymbol{S}}_k(t)^{\exp\left(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\right)-1}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t) \\ &= -\exp\big(f_k(\boldsymbol{\theta};\boldsymbol{x}_i)\big)\frac{\widehat{\mathbb{P}}(T > t|\boldsymbol{x}_i, Z = k))}{\widetilde{\boldsymbol{S}}_k(t)}\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)\end{aligned}$$

Here, $\frac{\partial}{\partial t}\widetilde{\boldsymbol{S}}_k(t)$ is the derivative of the baseline survival rate interpolated with a polynomial spline.

Appendix B. Censoring adjusted evaluation metrics

Area under ROC Curve (AUC): The ROC curve is defined as a plot between the True Positive Rate/Sensitivity (TPR) and the False Positive Rate (FPR) for all thresholds at which a classifier can be deployed. Note that the FPR is equal to 1−Specificity. We employ the technique proposed by Uno et al. (2007); Hung and Chiang (2010) to adjust the Sensitivity using IPCW estimates of the censoring distribution. The Specificity is computed on the uncensored instances.

$$\widehat{\text{Se}}(c, t) = \frac{\sum_{i=1}^{n} \omega_i \cdot \mathbb{1}\{\pi_i(t) > c, T_i \leq t\}\cdot}{\sum_{i=1}^{n} \omega_i \cdot \mathbb{1}\{T_i < t\}\cdot}; \quad \omega_i = \frac{\delta_i}{n \cdot \hat{G}(T_i)}; \quad \widehat{\text{Sp}}(c, t) = \frac{\sum_{i=1}^{n} \mathbb{1}\{\pi_i(t) \leq c, T_i > t\}\cdot}{\sum_{i=1}^{n} \mathbb{1}\{T_i > t\}\cdot}$$

$\widehat{\text{Se}}(c, t)$ and $\widehat{\text{Sp}}(c, t)$ refer to the estimated sensitivity and specificity at classification threshold c and time horizon t respectively. $\hat{G}(t)$ is a Kapaln-Meier estimator of the censoring distribution and $\pi_i(t)$ is the estimated survival probability, $\widehat{\mathbb{P}}(T > t|X = \boldsymbol{x}_i)$ by the classifier. This curve is plotted for all thresholds $c \in [0, 1]$ and the area under the curve is used to AUC. For a larger discussion around comparisons of various strategies to compute ROC curves in the presence of censoring refer to Kamarudin et al. (2017).

Time Dependent Concordance Index (C^{td}): Concordance Index estimates ranking ability by exhaustively comparing relative risks across all pairs of individuals in the test set within a fixed horizon of time.

$$C^{td}(t) = \mathbb{P}\big(\pi_i(t) \leq \pi_j(t)|\delta_i = 1, T_i < T_j, T_i \leq t\big)$$

Here, $\pi_i(t)$ is the estimated survival probability; T represent the event times. In order to deal with censoring we employ the censoring adjusted estimator for C^{td} that exploits IPCW estimates from a Kaplan-Meier estimate of the censoring distribution. The details are beyond the scope of this discussion and can be found in Uno et al. (2011) and Gerds et al. (2013).

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $R(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\text{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|R(t)) - R(t)\big|\big]$$

by partitioning the risk scores R into q quantiles $\{[r_j, r_{j+1})\}_{j=1}^{q}$. and computing the Kaplan-Meier estimate of the event rate $\text{KM}_j(t) \approx P(T > t|R \in [r_j, r_{j+1}))$, and the average risk score $\overline{R}_j = \frac{q}{n}\sum_{i:R_i \in [r_j, r_{j+1})} R_i$ in each bin. Altogether, the estimated ECE is

$$\widehat{\text{ECE}}(t) = \frac{1}{q}\sum_{j=1}^{q} |\text{KM}_j(t) - \overline{R}_j(t)|.$$

In practice, we fix the number of quantiles to be 20 for our experiments.

Brier Score (BS): The Brier Score involves computing the Mean Squared Error around the binary forecast of survival at a certain event quantile of interest. Brier Score is a proper scoring rule and can be decomposed into components that measure both discriminative performance and calibration.

$$\begin{aligned}
\text{BS}(t) &= \mathbb{E}_{\mathscr{D}}\big[\big(\mathbb{1}\{T_i > t\} - \widehat{\mathbb{P}}(T > t|X)\big)^2\big] \\
\widehat{\text{BS}}_{\text{IPCW}}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\pi_i(t)^2\mathbb{1}\{T \le t, \delta_i = 1\}}{\hat{G}_i(T_i)} + \frac{\big(1-\pi_i(t)\big)^2\mathbb{1}\{T > t\}}{\hat{G}_i(t)}\right]; \\
\text{where, } \pi_i(t) &= \widehat{\mathbb{P}}(T > t|X_i)
\end{aligned}$$

The adjusted Brier Score adjusted for Censoring using IPCW is given by $\widehat{\text{BS}}_{\text{IPCW}}(t)$ as proposed in (Graf et al., 1999; Gerds and Schumacher, 2006) Here, $\hat{G}(.)$ is the Kaplan Meier estimate of the Censoring Distribution. When the Censoring distribution is independent of the Event distribution, the above quantity is an unbiased estimate of the Brier Score.

Appendix C. Hyper-Parameter tuning for the Baselines

In this section we specify the hyper parameter choices along with a short description over which we perform grid search for the baselines.

Deep Survival Machines (DSM): The choice of hyper parameters for DSM include the number of underlying survival distributions (k) the choice of each outcome survival distribution, the number of hidden layers and neurons for the representation learning network and the activations. We also tune the learning rate and batch size. The choices of hyperparam values is given in Table 2.

Deep Hit (DHT): For Deep Hit, we tune the the Number of Hidden Layers, dimensionality of the hidden layers and the activation function. We also tune the learning rate and minibatch size. Note that Deep Hit requires grid discretization of the output event time space. For the SUPPORT and FLCHAIN datasets we discretize the output grid by dividing it into bins of $\max(T)$ bins. Since, the SEER is a discrete event time dataset we divide the output grid for

Table 2: DSM Hyper-parameter Grid

Hyper-parameter	Grid
Outcome Distribution	{ 'Weibull' }
No. Clusters (k)	{ '3', '4' }
No. of Hidden Layers	{ '0', '1', '2' }
Hidden Layer Dim.	{ '50', '100' }
Batch Size	{ '128', '256' }
Learning Rate	{ '1e-4', '1e-3' }
Activation	{ 'SeLU' }

Table 3: DHT and FSN Hyper-parameter Grid

Hyper-parameter	Grid
No. of Hidden Layers	{ '1', '2' }
Hidden Layer Dim.	{ '50', '100' }
Batch Size	{ '128', '256' }
Learning Rate	{ '1e-4', '1e-3' }
Activation	{ 'ReLU' }

Deep Hit into $^{\max(T)}/_{10}$ bins.

Faraggi-Simon Net (FSN)/DeepSurv: Similar to Deep Hit, for FSN we tune the the Number of Hidden Layers, dimensionality of the hidden layers and the activation function. We also tune the learning rate and minibatch size.

Both FSN and DHT were implemented using the `pycox` (Kvamme et al., 2019) python package. Table 3 describes the hyper-parameter choices for both DHT and FSN.

Table 4: RSF Hyper-parameter Grid

Hyper-parameter	Grid
Max Depth	{ '5' }
No. of Trees	{'50' }
mtry	{'sqrt' ,50 , 75, 'all' }
min_node_split	{'150' , '200', '250' }

Random Survival Forest (RSF): For the RSF model we tune the number of trees and the maximum depth of each tree using the implementation as paart of the `pysurvival` Python package (Fotso et al., 19). Table 4 presents the chosen grid parameters.

Table 5: AFT and CPH Hyper-parameter Grid

Hyper-parameter	Grid
$\ell 2$ Penalty	{ '1e-3', '1e-2', '1e-1' }

For **Cox Proportional Hazards (CPH)** and **Accelerated Failure Time (AFT)** the only hyperamaeter is the $\ell 2$ penalty on the parameters. The grid choice is presented in Table 5.

Appendix D. Additional Results

D.1. Tabulated Results

In this section we present tabulated results for our experiments for the entire population and the minority demogrpahic on the three datasets.

Tables 6 and 7 present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the FLCHAIN dataset, respectively.

Tables 8 and 9 present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the SUPPORT dataset, respectively.

Tables 10 and 11 and present the C^{td}, AUC, ECE and Brier Score for the Entire Population and Minority Demographic on the SEER dataset, respectively.

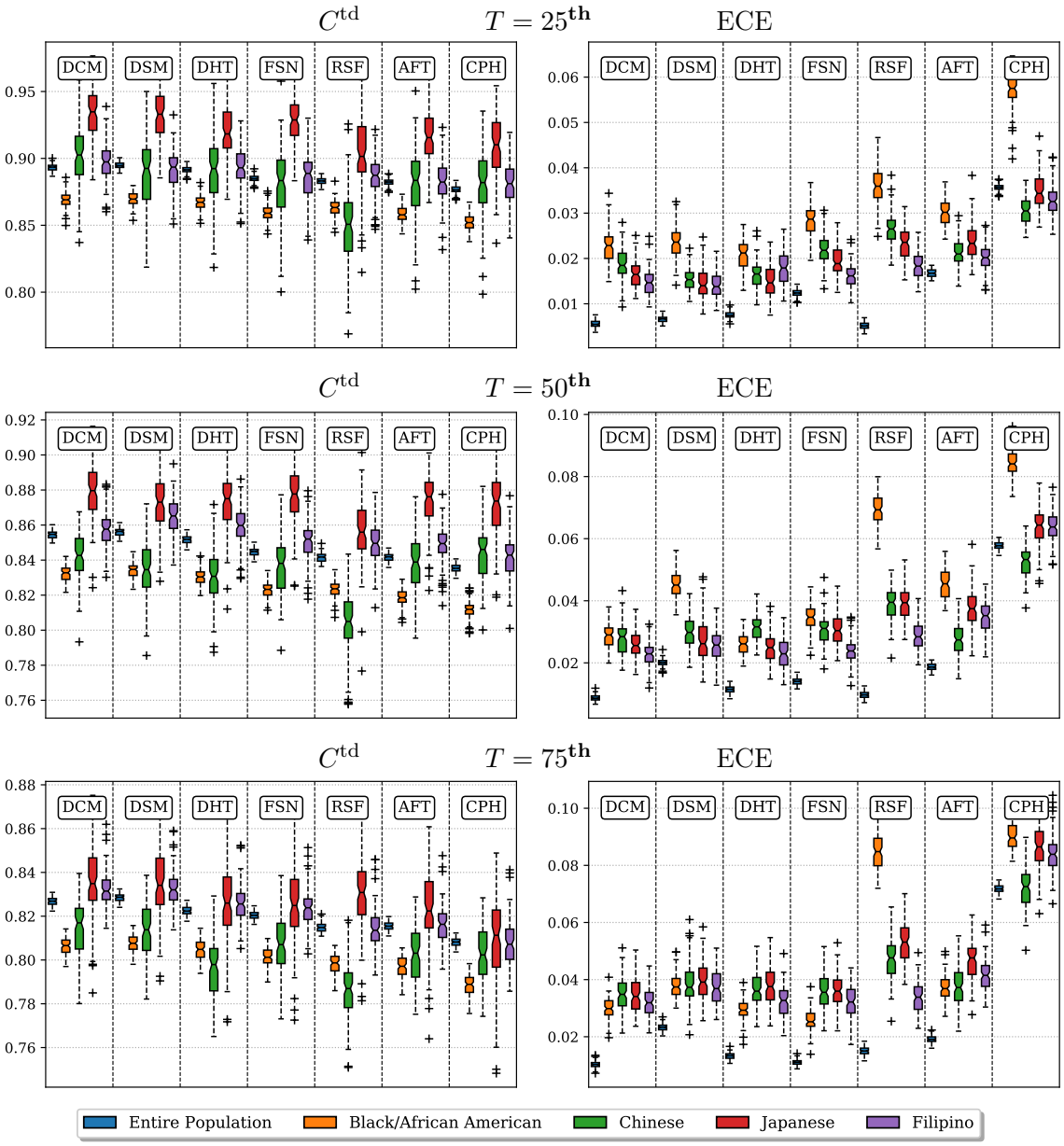


Figure 7: C^{td} (higher means better discrimination) and ECE (lower means better calibration) of proposed approach versus baselines at different quantiles of event times for the minority demographics. The rows represents different quantiles at which we evaluate the individual metrics. (Minorities in the dataset are denoted by different colors in the legend)

SEER has multiple minority classes. In Figure 7 we break results down by the top four largest minorities in the subset of SEER we are working with, 'Black/African American', 'Chinese', 'Japanese' and 'Filipino'.

D.2. Unawareness to Group Membership

In Table 12 we attempt to see how unawareness to the demographic affects the performance of Deep Cox Mixtures in terms of both, Calibration and Discrimination.

D.3. Dynamics of the Proposed MCMC EM Algorithm

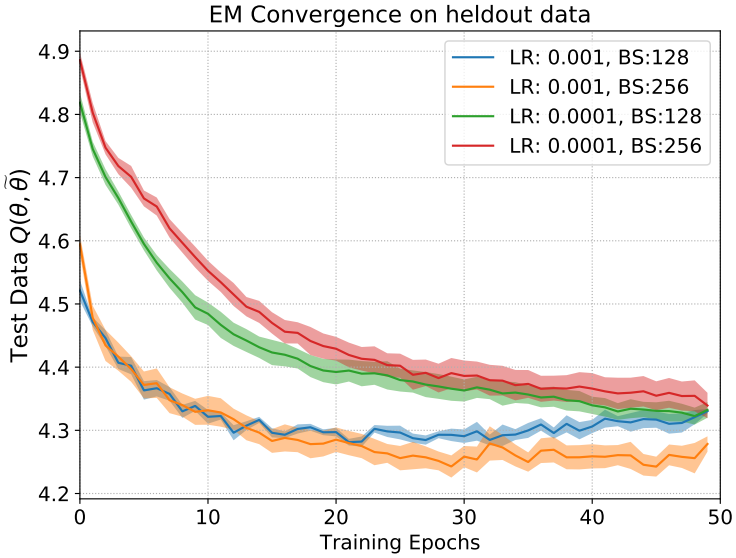


Figure 8: The estimated $Q(\theta, \widetilde{\theta})$ on a heldout set from the SUPPORT dataset for different hyper-parameters (LR: Learning Rate, BS: Batch size).

Figure 8 presents the estimated $Q(\theta, \widetilde{\theta})$ funciton for healdout dataset for the SUPPORT dataset. Empirically our proposed monte carlo EM monotonically decreases the $Q(\theta, \widetilde{\theta})$ suggesting good learning dynamics.

Supplementary References

Fotso, S. et al. (2019–). PySurvival: Open source package for survival analysis modeling.

Gerds, T. A., Kattan, M. W., Schumacher, M., and Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545.

Hung, H. and Chiang, C.-t. (2010). Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):53.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Uno, H., Cai, T., Tian, L., and Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6621 ± 0.0143	0.6696 ± 0.0110	0.6621 ± 0.0087
AFT	0.7914 ± 0.0107	0.7938 ± 0.0080	0.7911 ± 0.0060
RSF	0.7898 ± 0.0102	0.7908 ± 0.0078	0.7880 ± 0.0059
FSN	0.6353 ± 0.0146	0.6519 ± 0.0104	0.6608 ± 0.0081
DSM	0.8008 ± 0.0100	0.7988 ± 0.0078	0.7937 ± 0.0061
DHT	0.7669 ± 0.0104	0.7666 ± 0.0078	0.7636 ± 0.0059
DCM	0.7991 ± 0.0103	0.7988 ± 0.0077	0.7943 ± 0.0060

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6680 ± 0.0149	0.6827 ± 0.0120	0.6821 ± 0.0094
AFT	0.8032 ± 0.0110	0.8170 ± 0.0085	0.8257 ± 0.0063
RSF	0.8015 ± 0.0105	0.8142 ± 0.0083	0.8235 ± 0.0064
FSN	0.6416 ± 0.0150	0.6673 ± 0.0109	0.6904 ± 0.0090
DSM	0.8124 ± 0.0102	0.8218 ± 0.0083	0.8283 ± 0.0066
DHT	0.7771 ± 0.0106	0.7878 ± 0.0082	0.7936 ± 0.0064
DCM	0.8107 ± 0.0106	0.8219 ± 0.0082	0.8291 ± 0.0065

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0386 ± 0.0031	0.0699 ± 0.0042	0.0992 ± 0.0044
AFT	0.0141 ± 0.0024	0.0216 ± 0.0034	0.0212 ± 0.0034
RSF	0.0155 ± 0.0022	0.0198 ± 0.0027	0.0215 ± 0.0037
FSN	0.0214 ± 0.0027	0.0334 ± 0.0035	0.0381 ± 0.0046
DSM	0.0144 ± 0.0025	0.0214 ± 0.0030	0.0223 ± 0.0029
DHT	0.0283 ± 0.0029	0.0410 ± 0.0036	0.0505 ± 0.0041
DCM	0.0122 ± 0.0024	0.0169 ± 0.0033	0.0200 ± 0.0034

BS(t)(↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0671 ± 0.0027	0.1211 ± 0.0035	0.1665 ± 0.0037
AFT	0.0584 ± 0.0023	0.0991 ± 0.0028	0.1244 ± 0.0025
RSF	0.0603 ± 0.0023	0.1004 ± 0.0027	0.1250 ± 0.0026
FSN	0.0672 ± 0.0026	0.1199 ± 0.0029	0.1589 ± 0.0027
DSM	0.0578 ± 0.0022	0.0975 ± 0.0028	0.1224 ± 0.0026
DHT	0.0631 ± 0.0022	0.1086 ± 0.0026	0.1399 ± 0.0024
DCM	0.0582 ± 0.0023	0.0979 ± 0.0028	0.1228 ± 0.0026

Table 6: Results for various performance metrics on FLCHAIN (entire population) along with bootstrapped std errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6444 ± 0.0193	0.6692 ± 0.0160	0.6737 ± 0.0124
AFT	0.7822 ± 0.0158	0.7838 ± 0.0112	0.7875 ± 0.0087
RSF	0.7796 ± 0.0147	0.7799 ± 0.0113	0.7830 ± 0.0089
FSN	0.5746 ± 0.0211	0.6014 ± 0.0156	0.6212 ± 0.0131
DSM	0.7849 ± 0.0153	0.7886 ± 0.0113	0.7909 ± 0.0087
DHT	0.7607 ± 0.0153	0.7610 ± 0.0116	0.7631 ± 0.0092
DCM	0.7873 ± 0.0164	0.7893 ± 0.0116	0.7911 ± 0.0091

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6492 ± 0.0202	0.6842 ± 0.0175	0.6983 ± 0.0136
AFT	0.7944 ± 0.0163	0.8069 ± 0.0122	0.8230 ± 0.0095
RSF	0.7918 ± 0.0152	0.8028 ± 0.0124	0.8189 ± 0.0099
FSN	0.5774 ± 0.0219	0.6115 ± 0.0164	0.6477 ± 0.0148
DSM	0.7966 ± 0.0158	0.8118 ± 0.0123	0.8259 ± 0.0095
DHT	0.7710 ± 0.0157	0.7822 ± 0.0127	0.7938 ± 0.0104
DCM	0.7991 ± 0.0169	0.8122 ± 0.0126	0.8265 ± 0.0100

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0378 ± 0.0044	0.0642 ± 0.0056	0.0878 ± 0.0071
AFT	0.0221 ± 0.0035	0.0289 ± 0.0045	0.0329 ± 0.0046
RSF	0.0220 ± 0.0036	0.0330 ± 0.0046	0.0368 ± 0.0053
FSN	0.0325 ± 0.0043	0.0416 ± 0.0059	0.0545 ± 0.0068
DSM	0.0243 ± 0.0038	0.0323 ± 0.0048	0.0347 ± 0.0056
DHT	0.0328 ± 0.0037	0.0411 ± 0.0051	0.0525 ± 0.0056
DCM	0.0209 ± 0.0035	0.0298 ± 0.0054	0.0294 ± 0.0049

BS(t)(↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0693 ± 0.0043	0.1223 ± 0.0053	0.1626 ± 0.0057
AFT	0.0613 ± 0.0035	0.1031 ± 0.0040	0.1262 ± 0.0037
RSF	0.0624 ± 0.0036	0.1041 ± 0.0041	0.1273 ± 0.0038
FSN	0.0715 ± 0.0043	0.1278 ± 0.0051	0.1673 ± 0.0050
DSM	0.0609 ± 0.0035	0.1015 ± 0.0041	0.1244 ± 0.0037
DHT	0.0648 ± 0.0035	0.1107 ± 0.0038	0.1394 ± 0.0039
DCM	0.0607 ± 0.0035	0.1021 ± 0.0042	0.1249 ± 0.0039

Table 7: Results for various performance metrics on FLCHAIN (minority) along with bootstrapped std errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.6899 ± 0.0057	0.6713 ± 0.0040	0.6686 ± 0.0034
AFT	0.6826 ± 0.0057	0.6662 ± 0.0040	0.6657 ± 0.0034
RSF	0.7513 ± 0.0063	0.7104 ± 0.0045	0.6751 ± 0.0040
FSN	0.6988 ± 0.0059	0.6779 ± 0.0044	0.6736 ± 0.0037
DSM	0.7459 ± 0.0059	0.7042 ± 0.0038	0.6718 ± 0.0033
DHT	0.7302 ± 0.0067	0.6871 ± 0.0043	0.6575 ± 0.0038
DCM	0.7425 ± 0.0059	0.7057 ± 0.0042	0.6753 ± 0.0036

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.7011 ± 0.0061	0.6990 ± 0.0049	0.7214 ± 0.0049
AFT	0.6936 ± 0.0061	0.6943 ± 0.0049	0.7209 ± 0.0049
RSF	0.7663 ± 0.0066	0.7379 ± 0.0054	0.7273 ± 0.0054
FSN	0.7091 ± 0.0062	0.7050 ± 0.0052	0.7249 ± 0.0050
DSM	0.7606 ± 0.0063	0.7337 ± 0.0047	0.7236 ± 0.0050
DHT	0.7421 ± 0.0070	0.7123 ± 0.0052	0.7042 ± 0.0052
DCM	0.7576 ± 0.0065	0.7347 ± 0.0049	0.7256 ± 0.0054

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0201 ± 0.0029	0.0265 ± 0.0038	0.0310 ± 0.0041
AFT	0.0281 ± 0.0031	0.0617 ± 0.0048	0.0402 ± 0.0046
RSF	0.0241 ± 0.0032	0.0368 ± 0.0044	0.0348 ± 0.0041
FSN	0.0220 ± 0.0029	0.0267 ± 0.0036	0.0262 ± 0.0040
DSM	0.0341 ± 0.0033	0.0621 ± 0.0043	0.0315 ± 0.0047
DHT	0.0220 ± 0.0026	0.0351 ± 0.0037	0.0457 ± 0.0044
DCM	0.0179 ± 0.0030	0.0268 ± 0.0038	0.0256 ± 0.0037

BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.1334 ± 0.0023	0.1995 ± 0.0019	0.2136 ± 0.0016
AFT	0.1354 ± 0.0025	0.2051 ± 0.0023	0.2147 ± 0.0016
RSF	0.1240 ± 0.0023	0.1899 ± 0.0018	0.2109 ± 0.0017
FSN	0.1315 ± 0.0023	0.1981 ± 0.0020	0.2122 ± 0.0018
DSM	0.1271 ± 0.0024	0.1955 ± 0.0022	0.2130 ± 0.0017
DHT	0.1271 ± 0.0024	0.1971 ± 0.0016	0.2206 ± 0.0014
DCM	0.1258 ± 0.0024	0.1905 ± 0.0020	0.2118 ± 0.0019

Table 8: Results for various performance metrics on SUPPORT (entire population) along with bootstrapped std. errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.7161 ± 0.0126	0.6982 ± 0.0089	0.6905 ± 0.0078
AFT	0.7101 ± 0.0126	0.6941 ± 0.0089	0.6883 ± 0.0078
RSF	0.7503 ± 0.0120	0.7198 ± 0.0084	0.6974 ± 0.0084
FSN	0.7203 ± 0.0129	0.7025 ± 0.0090	0.6961 ± 0.0074
DSM	0.7548 ± 0.0132	0.7220 ± 0.0093	0.6939 ± 0.0079
DHT	0.7321 ± 0.0145	0.6943 ± 0.0099	0.6680 ± 0.0088
DCM	0.7570 ± 0.0130	0.7234 ± 0.0089	0.6939 ± 0.0079

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.7261 ± 0.0127	0.7348 ± 0.0109	0.7446 ± 0.0107
AFT	0.7199 ± 0.0127	0.7311 ± 0.0109	0.7446 ± 0.0108
RSF	0.7667 ± 0.0121	0.7536 ± 0.0106	0.7522 ± 0.0122
FSN	0.7283 ± 0.0128	0.7375 ± 0.0110	0.7518 ± 0.0101
DSM	0.7690 ± 0.0130	0.7594 ± 0.0113	0.7478 ± 0.0109
DHT	0.7400 ± 0.0143	0.7265 ± 0.0123	0.7129 ± 0.0120
DCM	0.7701 ± 0.0129	0.7588 ± 0.0109	0.7424 ± 0.0113

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0473 ± 0.0071	0.0610 ± 0.0084	0.0685 ± 0.0079
AFT	0.0530 ± 0.0075	0.0891 ± 0.0091	0.0741 ± 0.0085
RSF	0.0401 ± 0.0064	0.0608 ± 0.0077	0.0603 ± 0.0080
FSN	0.0418 ± 0.0067	0.0579 ± 0.0090	0.0601 ± 0.0097
DSM	0.0506 ± 0.0070	0.0818 ± 0.0094	0.0650 ± 0.0087
DHT	0.0483 ± 0.0070	0.0635 ± 0.0087	0.0696 ± 0.0089
DCM	0.0397 ± 0.0059	0.0550 ± 0.0080	0.0561 ± 0.0085

BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.1340 ± 0.0050	0.1943 ± 0.0042	0.2069 ± 0.0037
AFT	0.1363 ± 0.0054	0.2026 ± 0.0049	0.2090 ± 0.0039
RSF	0.1263 ± 0.0048	0.1870 ± 0.0039	0.2031 ± 0.0039
FSN	0.1319 ± 0.0051	0.1934 ± 0.0044	0.2037 ± 0.0040
DSM	0.1275 ± 0.0050	0.1919 ± 0.0047	0.2056 ± 0.0040
DHT	0.1298 ± 0.0049	0.1963 ± 0.0039	0.2186 ± 0.0036
DCM	0.1261 ± 0.0048	0.1868 ± 0.0044	0.2073 ± 0.0044

Table 9: Results for various performance metrics on SUPPORT (minority) along with bootstrapped std. errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8766 ± 0.0027	0.8354 ± 0.0024	0.8082 ± 0.0020
AFT	0.8823 ± 0.0026	0.8416 ± 0.0024	0.8155 ± 0.0020
RSF	0.8838 ± 0.0025	0.8421 ± 0.0025	0.8153 ± 0.0021
FSN	0.8850 ± 0.0025	0.8447 ± 0.0023	0.8204 ± 0.0019
DHT	0.8915 ± 0.0024	0.8517 ± 0.0024	0.8224 ± 0.0020
DSM	0.8949 ± 0.0022	0.8559 ± 0.0022	0.8281 ± 0.0019
DCM	0.8933 ± 0.0024	0.8550 ± 0.0022	0.8270 ± 0.0019

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8828 ± 0.0028	0.8526 ± 0.0025	0.8337 ± 0.0022
AFT	0.8893 ± 0.0026	0.8596 ± 0.0025	0.8424 ± 0.0021
RSF	0.8899 ± 0.0026	0.8594 ± 0.0026	0.8416 ± 0.0023
FSN	0.8921 ± 0.0026	0.8632 ± 0.0024	0.8477 ± 0.0021
DHT	0.8983 ± 0.0025	0.8701 ± 0.0025	0.8495 ± 0.0022
DSM	0.9022 ± 0.0023	0.8748 ± 0.0023	0.8566 ± 0.0020
DCM	0.9002 ± 0.0025	0.87350 ± 0.0023	0.8552 ± 0.0020

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0356 ± 0.0008	0.0577 ± 0.0012	0.0718 ± 0.0015
AFT	0.0168 ± 0.0008	0.0187 ± 0.0011	0.0192 ± 0.0011
RSF	0.0052 ± 0.0007	0.0092 ± 0.0010	0.0147 ± 0.0013
FSN	0.0124 ± 0.0008	0.0140 ± 0.0011	0.0111 ± 0.0011
DHT	0.0076 ± 0.0008	0.0115 ± 0.0011	0.0133 ± 0.0012
DSM	0.0067 ± 0.0007	0.0211 ± 0.0012	0.0259 ± 0.0014
DCM	0.0055 ± 0.0008	0.0087 ± 0.0010	0.0103 ± 0.0011

BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0501 ± 0.0007	0.0887 ± 0.0009	0.1206 ± 0.0009
AFT	0.0470 ± 0.0006	0.0827 ± 0.0009	0.1107 ± 0.0009
RSF	0.0447 ± 0.0006	0.0802 ± 0.0008	0.1095 ± 0.0010
FSN	0.0462 ± 0.0006	0.0800 ± 0.0008	0.1075 ± 0.0009
DHT	0.0450 ± 0.0006	0.0788 ± 0.0008	0.1074 ± 0.0010
DSM	0.0451 ± 0.0006	0.0797 ± 0.0008	0.1073 ± 0.0009
DCM	0.0450 ± 0.0006	0.0785 ± 0.0008	0.1064 ± 0.0010

Table 10: Results for various performance metrics on SEER (entire population) along with bootstrapped standard errors.

$C^{\text{td}}(t)$ (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8804 ± 0.0043	0.8405 ± 0.0039	0.8121 ± 0.0037
AFT	0.8865 ± 0.0042	0.8466 ± 0.0036	0.8204 ± 0.0035
RSF	0.8797 ± 0.0048	0.8379 ± 0.0038	0.8105 ± 0.0035
FSN	0.8870 ± 0.0043	0.8490 ± 0.0038	0.8248 ± 0.0036
DHT	0.8920 ± 0.0039	0.8540 ± 0.0038	0.8255 ± 0.0037
DSM	0.8908 ± 0.0038	0.8506 ± 0.0038	0.8243 ± 0.0036
DCM	0.8933 ± 0.0037	0.8558 ± 0.0036	0.8296 ± 0.0034

AUC(t) (↑)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.8888 ± 0.0043	0.8604 ± 0.0042	0.8398 ± 0.0042
AFT	0.8952 ± 0.0042	0.8676 ± 0.0039	0.8491 ± 0.0040
RSF	0.8867 ± 0.0048	0.8571 ± 0.0041	0.8373 ± 0.0039
FSN	0.8963 ± 0.0043	0.8702 ± 0.0040	0.8538 ± 0.0041
DHT	0.9002 ± 0.0039	0.8754 ± 0.0041	0.8540 ± 0.0041
DSM	0.9033 ± 0.0036	0.8770 ± 0.0039	0.8591 ± 0.0037
DCM	0.9020 ± 0.0037	0.8775 ± 0.0038	0.8595 ± 0.0038

ECE(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0399 ± 0.0018	0.0642 ± 0.0021	0.0764 ± 0.0028
AFT	0.0173 ± 0.0016	0.0271 ± 0.0022	0.0278 ± 0.0029
RSF	0.0112 ± 0.0016	0.0219 ± 0.0023	0.0270 ± 0.0029
FSN	0.0152 ± 0.0016	0.0198 ± 0.0025	0.0196 ± 0.0029
DHT	0.0107 ± 0.0015	0.0134 ± 0.0020	0.0170 ± 0.0024
DSM	0.0125 ± 0.0016	0.0292 ± 0.0023	0.0311 ± 0.0031
DCM	0.0105 ± 0.0016	0.0145 ± 0.0024	0.0169 ± 0.0024

BS(t) (↓)

Model	Quantiles		
	$t = 25$th	$t = 50$th	$t = 75$th
CPH	0.0563 ± 0.0014	0.0989 ± 0.0019	0.1285 ± 0.0020
AFT	0.0522 ± 0.0013	0.0907 ± 0.0019	0.1168 ± 0.0021
RSF	0.0508 ± 0.0014	0.0899 ± 0.0018	0.1190 ± 0.0021
FSN	0.0515 ± 0.0013	0.0877 ± 0.0017	0.1133 ± 0.0020
DHT	0.0509 ± 0.0012	0.0861 ± 0.0017	0.1135 ± 0.0021
DSM	0.0509 ± 0.0013	0.0882 ± 0.0018	0.1140 ± 0.0020
DCM	0.0508 ± 0.0012	0.0862 ± 0.0017	0.1127 ± 0.0020

Table 11: Results for various performance metrics on SEER (minority) along with bootstrapped standard errors.

$C^{\text{td}}(t)$ (↑)

Demographic	**Quantiles**		
	$t = 25$th	$t = 50$th	$t = 75$th
Entire Population	- 1.05 %	-1.34 %	-1.36 %
Minority Group	- 1.29 %	-1.48 %	-1.73 %

AUC(t) (↑)

Demographic	**Quantiles**		
	$t = 25$th	$t = 50$th	$t = 75$th
Entire Population	- 1.07 %	- 1.42 %	- 1.53 %
Minority Group	- 1.36 %	- 1.59 %	- 1.95 %

ECE(t) (↓)

Demographic	**Quantiles**		
	$t = 25$th	$t = 50$th	$t = 75$th
Entire Population	+ 17.30 %	+ 13.12 %	+ 10.74 %
Minority Group	- 0.80 %	+ 19.85 %	+ 20.99 %

BS(t) (↓)

Demographic	**Quantiles**		
	$t = 25$th	$t = 50$th	$t = 75$th
Entire Population	+ 1.89 %	+ 2.80 %	+ 3.13 %
Minority Group	+ 2.07 %	+ 4.53 %	+ 4.70 %

Table 12: Relative change in performance of Deep Cox Mixtures on the SEER dataset for the Entire Population and the Minority Demographic when unaware of the protected group membership. Overall, the performance in terms of both Discrimination and Calibration drops when DCM is made unaware of the protected groups. The relative deterioration in performance is worse for the minority demographic, suggesting unawareness to protected attribute being harmful in terms of the above performance metrics.