Proceedings of Machine Learning Research 219:1–45, 2023 Machine Learning for Healthcare

Fair Survival Time Prediction via Mutual Information Minimization

Hyungrok Do HYUNGROK.DO@NYULANGONE.ORG
Department of Population Health
NYU Grossman School of Medicine

Yuxin Chang YUXINC20@UCI.EDU
Department of Computer Science
University of California, Irvine

Yoon Sang Cho YOONSANG.CHO@NYULANGONE.ORG
Department of Population Health
NYU Grossman School of Medicine

Padhraic Smyth SMYTH@ICS.UCI.EDU
Department of Computer Science
University of California, Irvine

Judy Zhong JUDY.ZHONG@NYULANGONE.ORG
Department of Population Health
NYU Grossman School of Medicine

Abstract

Survival analysis is a general framework for predicting the time until a specific event occurs, often in the presence of censoring. Although this framework is widely used in practice, few studies to date have considered fairness for time-to-event outcomes, despite recent significant advances in the algorithmic fairness literature more broadly. In this paper, we propose a framework to achieve demographic parity in survival analysis models by minimizing the mutual information between predicted time-to-event and sensitive attributes. We show that our approach effectively minimizes mutual information to encourage statistical independence of time-to-event predictions and sensitive attributes. Furthermore, we propose four types of disparity assessment metrics based on common survival analysis metrics. Through experiments on multiple benchmark datasets, we demonstrate that by minimizing the dependence between the prediction and the sensitive attributes, our method can systematically improve the fairness of survival predictions and is robust to censoring.

1. Introduction

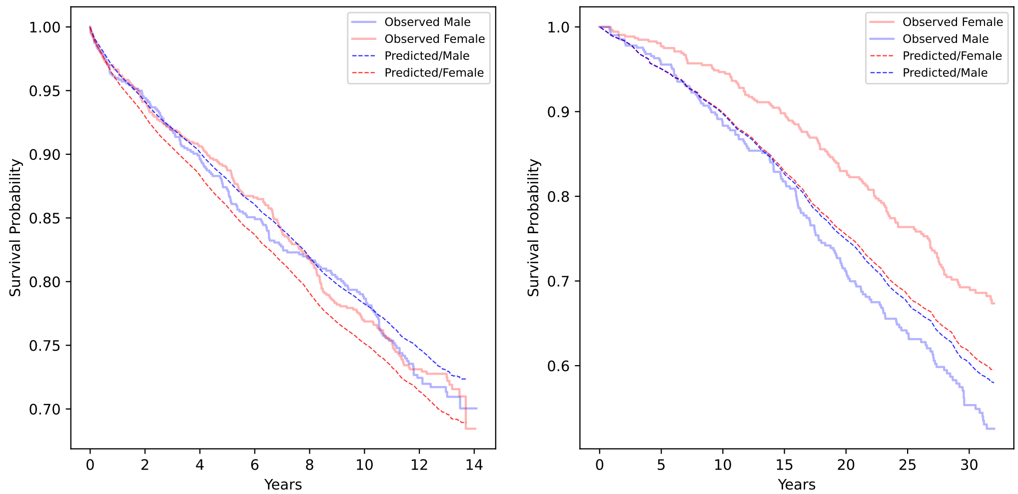
Though machine learning is increasingly being used to support and perform crucial decision-making tasks, recent research has clearly demonstrated that data-driven predictive models can often retain systematic biases that are present in the underlying data and can propagate these inequalities to their predictions. To address these issues, there has recently been a significant body of work in the machine learning community on algorithmic fairness in the context of predictive modeling, with the majority of this work focuses on classification and regression problems.

However, beyond classification and regression, there are problems in a broad range of areas, such as survival analysis, where the primary goal is to predict the time to an event of interest. Standard regression or classification models are typically inappropriate in such contexts due to *censoring*, where we have incomplete information about an individual's survival time when constructing a survival analysis model. This is particularly common in medical applications for example, where individuals are followed up for different lengths of time and some individuals have not had the event of interest occur at the end of the follow-up time.

Given the broad application of survival analysis in medical applications, we briefly discuss below a number of specific motivating examples (and related issues) from the medical field. Common medical applications of survival analysis include analyzing time to death or disease recurrence in clinical trials or predicting time to re-hospitalization using Electronic Health Records (EHRs). Unbalanced representations of subpopulation groups have been frequently reported in both contexts. For example, clinical trials are often biased and are not representative of racial/ethnic minority groups (Gianfrancesco et al., 2018). In addition, most EHRs of academic hospitals, minorities, and individuals with public insurance are under-represented with smaller sample sizes, shorter follow-up time, less encounters, and fewer lab measurements (Seyyed-Kalantari et al., 2020; Chen et al., 2021). There is an emerging recognition that such biases in data often lead to the unfair performance of predictive models (Paulus and Kent, 2020; Mhasawade et al., 2021; Gervasi et al., 2022) for life-saving decisions. As another example, when model-based predictions are used to prioritize patients for rationed services (e.g., organ transplantation, specialist referrals, or intensive care unit (ICU) services), prediction disparity can lead to systematically unfair treatments for the under-represented patients (Paulus and Kent, 2020). Most of these decision-making models rank candidates by their *predicted survival time*, such as ranking algorithms for organ transplantation (Nilsson et al., 2015) and ICU admission triage (Iwase et al., 2022). Therefore, the problem of ensuring fairness in healthcare-related predictive models is of vital importance, as these models are already being utilized to make highly sensitive and life-changing decisions.

However, there is a critical gap between the practical use of survival analysis and the development of fairness-aware methodologies in the research literature to achieve fair survival time prediction, because survival models often output the hazard function rather than the predicted survival time. Additionally, there has been limited discussion on effective fairness measurements for survival time predictions. This paper addresses these gaps by:

- Developing a general framework for Fair Survival Time Prediction (FAST) that *directly* achieves Demographic Parity (DP), between the predicted survival time $\mathcal{T}_\theta$ and sensitive attributes, rather than DP on model outputs, $f(\mathbf{X})$. This is critical in survival analysis because $\mathcal{T}_\theta$ takes inputs from $f(\mathbf{X})$ and others, including baseline survival probabilities affected by censoring C;



(a) FLChain (b) Framingham

Figure 1: Comparison of observed (Kaplan-Meier) and predicted survival curves for the FLChain and Framingham datasets by a non-fairness-aware model, separated by sensitive attributes. The predicted survival curves were generated using the DeepCox model, and both DNNSurv and DeepHit models showed similar patterns (see Appendix A). For both datasets, the non-fairness-aware model consistently overestimated the event risk for females, highlighting the presence of systematic bias and disparities in survival analysis. These results suggest that non-fairness-aware models may systematically produce unfair and biased outcomes across sensitive attributes in survival analysis.

- Proposing a series of comprehensive and interpretable fairness assessment metrics for survival time predictions, including concordance and calibration;
- Illustrating via empirical studies on multiple survival analysis datasets[1] that the proposed FAST method can systematically improve prediction parities while maintaining reasonable degrees of overall performances;
- Demonstrating two unique advantages of FAST methodology: (1) compatibility with a broad range of existing survival analysis models, and (2) robustness when the disparity originates from censoring, which is a specific (and common) issue in survival analysis.

Motivating Example: Presence of Prediction Disparity

To assess the presence of prediction disparities, we implemented a non-fairness-aware model, the Deep Cox proportional hazard model (DeepCox), on the FLChain and Framingham datasets, where the sensitive attribute considered was “sex” (female and male). Sex was not included as a covariate in the prediction model. We compared the observed survival curves

1. Code available at https://github.com/nyumed-judy-lab/fair-survival

obtained from the Kaplan-Meier method with the predicted survival curves obtained from DeepCox for each group in each dataset.

For FLChain, the prediction for the male group closely followed its observed curve, while the predicted curve for the female group was below the observed curve (Figure 1-(a)). These findings indicate that DeepCox systematically overestimates the event risk for females but performs relatively well for males. Similar results were observed for Framingham (Figure 1-(b)). We also tried other non-fairness-aware models, including DNNSurv and DeepHit models, and they both showed similar patterns (see Appendix A). Overall, these results highlight the presence of prediction disparities, which can lead to systematic bias and unfairness in the prediction outcomes for sensitive attributes in survival analysis.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our experience in assessing prediction disparities of survival analysis models and developing fair survival analysis models based on deep neural networks has led to a number of generalizable insights we discuss below. The primary insights are:

- In healthcare, it is important for evaluation metrics of machine learning methods to reflect their practical use if they were to be deployed. While it is common for researchers to use metrics from previous studies, it is also crucial to examine whether these metrics are realistic and clinically relevant.

- When assessing prediction disparity in a survival time prediction model, it is essential to consider various factors. For example, different forms of systematic prediction disparity, including concordance, calibration, and Brier score, may exist in a survival analysis model, as can be seen in our motivating example. It is important to note that the presence of prediction disparity in one type may not necessarily indicate its presence in other types.

- Researchers tend to overlook the importance of the baseline hazard function in the Cox proportional hazard model because of the proportional hazard assumption. However, when evaluating prediction disparity in survival analysis, it is crucial to take the baseline hazard function into account for calibration performance, such as expected ℓ_1 calibration error, and the Brier score.

2. Related Work

Broadly speaking, there are three main strategies that are pursued in the algorithmic fairness research literature: (i) data pre-processing (Calmon et al., 2017; Li and Liu, 2022), (ii) in-process approaches which enforce fairness during model training (Zafar et al., 2017; Donini et al., 2018; Agarwal et al., 2019; Kleindessner et al., 2022; Shah et al., 2022; Do et al., 2022), and (iii) post-processing to adjust a model's predictions to achieve fairness after the model training (Hardt et al., 2016; Wei et al., 2020; Chzhen et al., 2020; Soen et al., 2022). Our work fits in the in-processing approach. Early examples of this approach include fair representation learning (FRL), e.g., the work of Zemel et al. (2013) reduces DP for binary classification, with significant follow-up work for a variety of other fairness criteria and various downstream tasks (Madras et al., 2018; Roy and Boddeti, 2019; Gupta et al., 2021; Kim et al., 2022; Shui et al., 2022).

Nonetheless, although there has been considerable prior work on FRL and other approaches for fair prediction in contexts such as classification and regression, there has been relatively little work focusing on fair survival analysis specifically. In particular, because the output of a survival analysis model is typically a hazard function (namely the instantaneous rate of risk of an event at some time t), common definitions of fairness using DP are not straightforwardly applicable because they cannot be defined directly on the output of a survival model. A number of recent papers have attempted to address this issue in different ways. Zhang and Weiss (2021, 2022) proposed the fair survival random forest model (FSRF), using a fair splitting criterion based on the log-rank test statistic. However, their approach is limited to tree-based models and requires discretization of real-valued sensitive attributes. Keya et al. (2021) developed the fair Cox proportional hazard models and associated fairness metrics by equalizing the proportional hazards and ignoring the baseline hazards, but requiring restrictive parametric assumptions of the Cox proportional hazard model. Rahman and Purushotham (2022) proposed pseudo value-based survival models by adding a fairness penalty term defined on predicted survival probability at time t. A significant limitation of this approach is that it only applies to models with a pseudo value-based loss function. Finally, Hu and Chen (2022) applied distributionally robust optimization (DRO), originally developed for binary classification (Hashimoto et al., 2018), to Cox proportional hazard model to achieve fairness. Besides, Curth et al. (2021) studied learning heterogeneous treatment effects for survival analysis, which is closely related to fairness problems.

In summary, general techniques to impose fairness in classification and regression tasks are not directly applicable to survival analysis. In addition, the (limited) prior work on fair survival analysis requires strong model assumptions and/or the use of ad-hoc fairness assessment metrics that lack clear interpretation. Our framework overcomes these limitations by using interpretable assessment metrics and being flexible and general so it can be compatible with any survival model that has a differentiable likelihood. Our approach also bridges the gap between fairness definitions and the predicted survival time; we demonstrate the utility of this via a novel set of comprehensive evaluation metrics derived from widely-used metrics for survival analysis.

3. Problem Formulation

Survival analysis involves the estimation of the probability distribution of time-to-event in the presence of censoring. In a typical survival analysis setting, for each individual, we have realizations of covariates $\mathbf{X}$, a follow-up time O, and an event indicator Δ, which takes value 1 if the event precedes the censoring and 0 otherwise. In addition, we have K groups corresponding to K possible values in $\mathcal{A} = \{\alpha_1, \cdots, \alpha_K\}$ of a sensitive attribute A such as race/ethnicity or gender A. Thus, we have a dataset $\mathcal{D} = \{(\mathbf{x}_i, a_i, o_i, \delta_i) \in \mathcal{X} \times \mathcal{A} \times \mathbb{R}_+ \times \mathbb{B} : i = 1, \cdots, n\}$ where $\mathbf{x}_i$, a_i, o_i, and δ_i are realizations of $\mathbf{X}$, A, O, and Δ for individual i.

In this paper, we adopt the usual independent censoring assumption in the survival analysis. Let T and C be the event and censoring time, respectively, so that $O = \min\{T, C\}$ and $\Delta = \mathbb{I}(T \leq C)$.

Under the independent censoring assumption, we learn a model, parameterized by θ, by minimizing the negative log-likelihood: $\widehat{\theta}_{\mathrm{MLE}} = \mathrm{argmin}_{\theta} - \ell(\theta; \mathcal{D})$, where

$$\ell(\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \Big[\delta_i \log \lambda_\theta(o_i|\mathbf{x}_i) - \Lambda_\theta(o_i|\mathbf{x}_i) \Big], \tag{1}$$

and $\Lambda_\theta(\cdot|\mathbf{x})$ is the cumulative hazard function such that $\Lambda_\theta(t|\mathbf{x}) = -\log S_\theta(t|\mathbf{x})$, for all $t \geq 0$. $S_\theta(t|\mathbf{x}) = P(T > t|\mathbf{x})$ is the survival function given $\mathbf{x}$, and $\lambda_\theta(t|\mathbf{x})$ is the conditional hazard function defined as $\frac{d}{dt}\Lambda_\theta(t|\mathbf{x})$. The actual form of the hazard function depends on the choice of model.

This conventional MLE approach for survival analysis often results in dependence between the prediction and the sensitive attributes, which may lead to disparity of prediction performance among the sensitive groups. To mitigate the prediction disparity, we can add a penalty term $\mathcal{P}(\theta; \mathcal{D})$, yielding the following penalized negative log-likelihood problem:

$$\widehat{\theta}_{\mathrm{FAST}} = \underset{\theta}{\mathrm{argmin}} - \ell(\theta; \mathcal{D}) + \gamma \mathcal{P}(\theta; \mathcal{D}), \tag{2}$$

where $\gamma > 0$ is a hyperparameter. By jointly minimizing the negative log-likelihood for prediction as well as the fairness penalty term, we can obtain a survival model that has less prediction disparity compared to the conventional MLE approach. In the following section, we introduce how to formulate the fairness penalty term.

3.1. DP for Survival Analysis

One difference between the survival analysis and other prediction tasks is that a survival model produces the conditional hazard value conditioned on $\mathbf{x}$, while models for other tasks typically output the conditional expected outcome value given $\mathbf{x}$. We denote the predicted survival time as a random variable $\mathcal{T}_\theta$, distributed over $\mathbb{R}_+$, from a survival analysis model parameterized by θ. The cumulative probability distribution function of $\mathcal{T}_\theta$ is defined by the cumulative hazard function as $P(\mathcal{T}_\theta(\mathbf{x}) \leq t) = 1 - \exp(-\Lambda_\theta(t|\mathbf{x}))$.

Definition 1 (Demographic Parity (Kamiran and Calders, 2009)) *We say the survival analysis model, parameterized by θ, satisfies* demographic parity *if $\mathcal{T}_\theta$ is statistically independent of the sensitive attribute A, i.e., $A \perp\!\!\!\perp \mathcal{T}_\theta$. In other words, $\hat{S}_\theta(t|A = \alpha_j) = \hat{S}_\theta(t|A = \alpha_k)$ for any $t \in [0, \infty)$ and $\alpha_j, \alpha_k \in \mathcal{A}$.*

To achieve DP, we need to minimize the dependence between the sensitive attribute A and the predicted survival time $\mathcal{T}_\theta$. The dependence between two random variables can be quantified in several ways. For example, *mutual information* (MI), which quantifies the amount of information obtained about one random variable by observing the other one, captures non-linear statistical dependencies between random variables and is often considered as a measure of true statistical dependence (Kinney and Atwal, 2014). Some alternative choices for the dependence measure are discussed in Appendix B.1.

Definition 2 (Mutual Information of A and $\mathcal{T}_\theta$) *Mutual information that quantifies the dependence of the sensitive attribute A and the predicted survival time $\mathcal{T}_\theta$ is defined as*

$$\mathcal{I}(A, \mathcal{T}_\theta) = \int_{\mathcal{A}\times\mathbb{R}} \log \frac{d\mathbb{P}_{A\mathcal{T}_\theta}}{d\mathbb{P}_A \otimes d\mathbb{P}_{\mathcal{T}_\theta}} d\mathbb{P}_{A\mathcal{T}_\theta}, \tag{3}$$

where $\mathbb{P}_{A\mathcal{T}_\theta}$ is the joint CDF, $\mathbb{P}_A$ and $\mathbb{P}_{\mathcal{T}_\theta}$ are the marginal CDFs of A and $\mathcal{T}_\theta$ respectively. MI is nonnegative, and a smaller value implies a weaker dependence between A and $\mathcal{T}_\theta$. Moreover, $\mathcal{I}(A, \mathcal{T}_\theta) = 0$ if and only if $A \perp\!\!\!\perp \mathcal{T}_\theta$.

We propose a **FA**ir **S**urvival **T**ime (FAST) prediction framework, as the minimizer of the weighted sum of the negative log-likelihood and the MI between A and $\mathcal{T}_\theta$:

$$\widehat{\theta}_{\text{FAST}} = \underset{\theta}{\operatorname{argmin}} - \ell(\theta; \mathcal{D}) + \gamma \mathcal{I}(A, \mathcal{T}_\theta). \tag{4}$$

We note that formulating the fair learning problem as MI minimization is prevalent in the literature (Creager et al., 2019; Song et al., 2019; Zhu et al., 2021; Grari et al., 2021; Zheng and Li, 2022), however, none of them can be applied to the survival analysis directly.

3.2. Mutual Information Estimation

Since the mutual information of A and $\mathcal{T}_\theta$ will often not have a closed form, we adopt a non-parametric approach, in particular, the Mutual Information Neural Estimation (MINE) proposed by Belghazi et al. (2018). This allows us to estimate the mutual information from data samples without assuming a distributional form of A or $\mathcal{T}_\theta$.

The main idea of MINE is to maximize the lower bound of the mutual information utilizing its dual representation (Belghazi et al., 2018). Let $\{a_i \in \mathcal{A} : i = 1, \cdots, n\}$ be a set of sensitive attributes and $\{\tau_i \sim \mathcal{T}_\theta(\mathbf{x}_i) : i = 1, \cdots, n\}$ be a set of samples drawn from $\mathcal{T}_\theta(\mathbf{x}_i)$, respectively. Let $\{a'_i \in \mathcal{A} : i = 1, \cdots, n\}$ be another set of sensitive attributes obtained by randomly rearranging the elements of $\{a_i \in \mathcal{A} : i = 1, \cdots, n\}$. The estimate of mutual information $\hat{\mathcal{I}}(A, \mathcal{T}_\theta; \mathcal{D})$ of A and $\mathcal{T}_\theta$ can be obtained through maximizing its lower bound $\tilde{\mathcal{I}}_\omega(A, \mathcal{T}_\theta; \mathcal{D})$

$$\hat{\mathcal{I}}(A, \mathcal{T}_\theta; \mathcal{D}) = \sup_\omega \tilde{\mathcal{I}}_\omega(A, \mathcal{T}_\theta; \mathcal{D}) = \sup_\omega \frac{1}{n} \sum_{i=1}^n \psi_\omega\big(a_i, \tau_i\big) - \log \left[\frac{1}{n} \sum_{i=1}^n \exp \psi_\omega\big(a'_i, \tau_i\big) \right], \tag{5}$$

where $\psi_\omega : (a_i, \tau_i) \mapsto \mathbb{R}$ is a function parameterized by a neural network with parameters ω. Note (a_i, τ_i) is a sample drawn from the joint distribution $\mathbb{P}_{A\mathcal{T}_\theta}$ and (a'_i, τ_i) is drawn from the product of the marginals $\mathbb{P}_A \otimes \mathbb{P}_{\mathcal{T}_\theta}$. Belghazi et al. (2018) proved that MINE is a strongly consistent estimator of mutual information and linearly scalable in both dimensionality and sample size. Moreover, it is trainable through backpropagation, so that we can plug it into Equation (7).

3.3. Log-likelihood

FAST does not require a specific form of log-likelihood and is compatible with any survival analysis model with a differentiable log-likelihood. If one chooses to use the Cox proportional hazard (PH) model, the log-likelihood is

$$\ell(\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \Big[\delta_i (\log \lambda_0(o_i) + f_\theta(\mathbf{x}_i) - \Lambda_0(o_i) \exp f_\theta(\mathbf{x}_i) \Big], \tag{6}$$

where λ_0 and Λ_0 are the baseline hazard function and the cumulative baseline hazard functions respectively, and $f_\theta : \mathcal{X} \to \mathbb{R}$ is the log proportional hazard function which is parameterized by parameters θ. We present several alternative choices for the log-likelihood function in Appendix C.

Algorithm 1 SGD for FAST

Input: Data $\mathcal{D} = \{(\mathbf{x}_i, a_i, o_i, \delta_i\}_{i=1}^n$, hyperparameter γ, learning rates η_θ, η_ω, batch size b.
Output: $\widehat{\theta}_{\text{FAST}}$ solving (7).
$\theta^{(0)}, \omega^{(0)} \leftarrow$ initialize
while not converged do
  Draw a minibatch of samples from $\mathcal{D}$: $\mathfrak{B} = \{(\mathbf{x}_i, a_i, o_i, \delta_i)\}_{i=1}^b$
  Draw time-to-event from $\mathcal{T}_{\theta^{(t)}}(\mathbf{x}_i)$: $\{\tau_i\}_{i=1}^b$
  Draw another minibatch of samples from A: $\{a_i'\}_{i=1}^b$
  Prepare a minibatch for $\tilde{\mathcal{I}}$: $\mathfrak{C} = \{(a_i, \tau_i, a_i')\}_{i=1}^b$
  Update θ by descending its stochastic gradient.
  $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta(\nabla \ell(\theta^{(t)}; \mathfrak{B}) - \gamma \nabla \tilde{\mathcal{I}}_\omega(A, \mathcal{T}_{\theta^{(t)}}; \mathfrak{C}))$
  Update ω by ascending its stochastic gradient.
  $\omega^{(t+1)} \leftarrow \omega^{(t)} + \eta_\omega \nabla \tilde{\mathcal{I}}_\omega(A, \mathcal{T}_{\theta^{(t)}}; \mathfrak{C})$
end while
$\widehat{\theta}_{\text{FAST}} = \theta^{(t)}$

3.4. Estimation of FAST

With Equation (5), the estimation of FAST becomes:

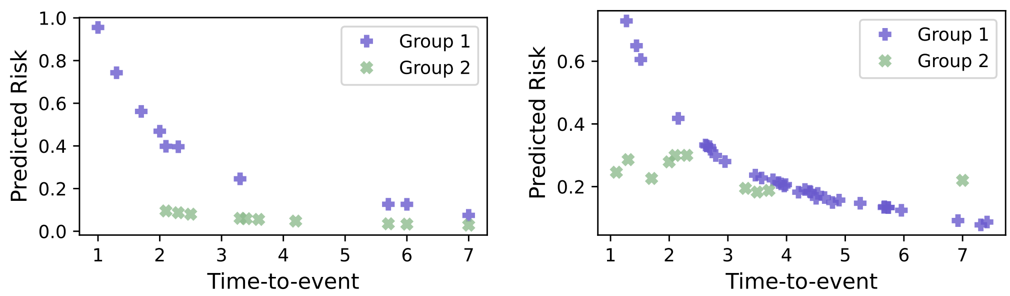
$$\min_\theta \max_\omega -\ell(\theta; \mathcal{D}) + \gamma \tilde{\mathcal{I}}_\omega(A, \mathcal{T}_\theta; \mathcal{D}). \tag{7}$$

We solve the minimax problem using minibatch stochastic gradient descent (SGD) optimization. Given each batch, we simultaneously update θ and ω by ascending and descending their gradients, as illustrated in Algorithm 1. At each iteration, the algorithm randomly draws minibatches of size b from $\mathbb{P}_{\mathbf{X}AO\Delta}$ and $\mathbb{P}_A$. Then, it generates random samples from $\mathcal{T}_\theta(\mathbf{x}_i)$ for each i. The exact method to sample τ_i differs by choice of model, but in most cases, we can draw samples using inverse transform sampling (see Appendix C for details). Then, we take random samples drawn from the joint distribution $\mathbb{P}_{A\mathcal{T}_\theta}$ and the product of marginal distributions $\mathbb{P}_A \otimes \mathbb{P}_{\mathcal{T}_\theta}$ and plug them into (5). This allows us to calculate the gradient of $\tilde{\mathcal{I}}_\omega(A, \mathcal{T}_\theta; \mathcal{D})$ with respect to ω and θ, respectively. As studied in Belghazi et al. (2018), the naive minibatch SGD gradients for $\tilde{\mathcal{I}}_\omega(A, \mathcal{T}_\theta; \mathcal{D})$ are biased, so we adopt the moving average technique to reduce the bias.

The gradient of ℓ with respect to θ can be straightforwardly computed. For the CoxPH log-likelihood function (6) with f_θ parameterized by a neural network, we first estimate the baseline hazard functions $\hat{\lambda}_0$ and $\hat{\Lambda}_0$. We then take the full log-likelihood approach to solve $\ell(\theta; \mathcal{D}, \hat{\lambda}_0, \hat{\Lambda}_0)$.

3.5. Choice of the Hyperparameters

We take a general strategy to select the hyperparameter, namely to minimize the disparity while placing constraints on the overall predictive performance. Specifically, we evaluate the performances and disparities of models trained with different γ values on a held-out validation set, and choose the γ with the smallest disparity while maintaining the overall performance within a pre-determined acceptable margin, such as $\pm 5\%$, of that under the vanilla model



(a) (b)

Figure 2: (Left) The case that a model discriminates the risks well within each group, however, significantly underestimates the risks for Group 2. Here, $\text{CI}_1^s = \text{CI}_2^s = 1.0$. However, $\text{CI}_1^p = 0.93$ and $\text{CI}_2^p = 0.64$. That is, ΔCI^s does not capture the disparity. (Right) The case that a model performs well on over-represented Group 1, however, provides poor risk estimations on under-represented Group 2. In this case, $\text{CI}_1^s = 0.95$ and $\text{CI}_2^s = 0.53$, while $\text{CI}_1^p = 0.89$ and $\text{CI}^p = 0.82$. That is, ΔCI^s reflects the disparity better.

with $\gamma = 0$. In order to investigate the sensitivity of metrics for hyperparameter-tuning purposes, we propose a series of evaluation metrics in Section 4, and compare their empirical performances in Section 5. We note that Keya et al. (2021) considered a similar strategy for selecting the fairness hyperparameter.

4. Fairness Assessment for Survival Analysis

The prediction performance of a survival model is usually assessed by concordance, discrimination, and calibration. There has been limited discussion to date in the survival analysis literature on what metrics are recommended to assess the disparity of prediction performances among sensitive attribute groups. One straightforward assessment is to calculate the average (absolute) differences of any metric between the sensitive attribute groups. In this section, we discuss that this is often insufficient, especially for concordance and discrimination metrics.

We will denote the results from a survival analysis model as $\{(o_i, r_i, \delta_i) : i = 1, \cdots, n\}$, where $o_i = \min\{c_i, t_i\}$ is the observed follow-up time, $r_i = P(\mathcal{T}_{\widehat{\theta}}(\mathbf{x}_i) < t)$ is the predicted risk at time t provided by a survival analysis model parameterized by $\widehat{\theta}$, and $\delta_i = \mathbb{I}(t_i \leq c_i)$ is the observed event indicator.

4.1. Concordance

The concordance index (C-index) is a widely used metric for evaluating survival analysis models (Harrell et al., 1982, 1984, 1996). It captures the rank correlation between the survival time and predicted risks – a survival model tends to assign high-risk individuals who have shorter survival times with greater C-index. C-index is calculated as the proportion of the number of concordant pairs out of the number of valid pairs $\text{CI} = |\mathcal{S}^c|/|\mathcal{S}^v|$, where $\mathcal{S}^v = \{(i,j) : o_i \leq o_j, \delta_i = 1\}$ is the set of valid pairs, and $\mathcal{S}^c = \{(i,j) \in \mathcal{S}^v : r_i > r_j\}$ is the set of concordant pairs.

However, to evaluate prediction concordance for subjects in a subgroup, there are two versions of the C-index, differing in how the valid and concordant pairs are defined. If one considers only the subjects in a sensitive attribute group k, the sets of valid pairs and concordant pairs are defined as $\mathcal{S}_k^{sv} = \{(i,j) : o_i \leq o_j, \delta_i = 1, a_i = a_j = \alpha_k\}$ and $\mathcal{S}_k^{sc} = \{(i,j) \in \mathcal{S}_k^{sv} : r_i > r_j\}$ respectively. We refer to this as the **stratified group C-index**, $\text{CI}_k^s = |\mathcal{S}_k^{sc}|/|\mathcal{S}_k^{sv}|$.

The stratified group C-index measures how a model correctly arranges the risks and times-to-event within the group. However, it cannot capture the disparity of group k relative to other groups. For instance, considering a model that underestimates one group's risk compared to another (Figure 2-(a)), its stratified C-indices for all groups can be good, even if the rank correlation between the observed survival time and predicted risks are relatively worse for the group.

This encourages us to consider another group-level C-index to reflect the macro-level assessment. We define the set of valid pairs and concordant pairs by $\mathcal{S}_k^{pv} = \{(i,j) : o_i \leq o_j, \delta_i = 1, a_i = \alpha_k\}$ and $\mathcal{S}_k^{pc} = \{(i,j) \in \mathcal{S}^{pv} : r_i > r_j\}$. Note that this definition pairs instances of a specific group with pooled candidates from the entire dataset. Thus, we refer to this as the **pooled group C-index**, $\text{CI}_k^p = |\mathcal{S}_k^{pc}|/|\mathcal{S}_k^{pv}|$.

In Figure 2-(a), CI_2^p is significantly worse than that of Group 1. However, in another situation when a model predicts like Figure 2-(b), CI_k^p cannot capture the within-group concordance, while the stratified group C-index can. Therefore, we argue that disparities of both versions of C-indices are needed to comprehensively measure the concordance disparity. We define the stratified group C-index disparity ΔCI^s, as the maximum of pairwise absolute differences of stratified group C-indices from all possible groups, and similarly for the pooled group C-index disparity ΔCI^p. Note that, ΔCI^p has also been considered in Zhang and Weiss (2021) as the "concordance imparity." Also, a similar comparison between the stratified and pooled group AUCs, in the context of binary classification, was discussed in Yang et al. (2023).

4.2. Calibration

Calibration is an important component for assessing survival time prediction models, which measures the agreement between predicted probabilities and observed event rates within a given duration of time. Different methods have been proposed to evaluate the calibration of survival models. They typically involve constructing a calibration curve by dividing individuals into subsets based on predicted event probabilities and comparing the mean predicted event probabilities with the observed event prevalence for each subset. Graphical representations are often used to compare the results with a diagonal line that represents perfect calibration. For more detailed information on constructing and summarizing calibration curves for survival analysis, refer to Austin et al. (2020); Haider et al. (2020); Goldstein et al. (2020) and references therein. In our study, we quantify the calibration performance of survival models using the expected ℓ_1 calibration error (ECE). The ECE is a popular metric for classification problems using deep neural networks (Guo et al., 2017), and it is

Table 1: Benchmark datasets and their statistics. p is the number of covariates, n is the number of instances, A is the sensitive attribute, and $\mathcal{C}$ is the percentage of censoring.

Datasets	Outcome	p	A	$n(\mathcal{C})$
FLChain (Kyle et al., 2006)	Time to Death	17	Female	4,347 (73.3)
			Male	3,524 (72.5)
Framingham (Mahmood et al., 2014)	Time to Death	5	Female	2,650 (75.5)
			Male	2,049 (59.8)
SUPPORT (Knaus et al., 1995)	Time to Death	66	White	7,191 (31.2)
			Black	1,391 (34.7)
			Hispanic	290 (40.3)
			Other	191 (27.2)

also used to evaluate the calibration performance of deep survival models (Nagpal et al., 2021). The ECE calculates the average absolute difference between the mean predicted event probabilities and observed event rates conditional on the predicted event probability. The ECE at time t is defined as:

$$\textsc{Ece} = \sum_{j=1}^{q} \frac{|\mathcal{Q}_j|}{n} \Big|(1 - \mathrm{KM}_j(t)) - \bar{r}_j\Big|, \tag{8}$$

Here, we divide the predicted risk scores into q quantiles (bins) $\mathcal{Q}_j = [r_j, r_{j+1})$ for $j = 1, \cdots, q$, estimate the Kaplan-Meier survival probabilities KM_j, and calculate the average risk score $\bar{r}_j = \frac{1}{|\mathcal{Q}_j|} \sum i \in \mathcal{Q}_j r_i$ for each bin $\mathcal{Q}_j$. We define the disparity of the ECE, ΔEce, as the maximum pairwise absolute difference of ECEs across all possible groups. Note that Zhang and Weiss (2022) considered hypothesis testing for fair calibration as a series of Hosmer-Lemeshow goodness-of-fit tests for each group.

4.3. Brier Score

The Brier score (Byers et al., 1951) is the most widely used metric to evaluate survival models' concordance and calibration performances (Murphy, 1972; DeGroot and Fienberg, 1983; Haider et al., 2020). It is defined as the average squared distance between the observed survival status and the predicted survival probability.

$$\textsc{Brier} = \frac{1}{n} \sum_{i=1}^{n} [(1 - r_i)^2 \mathbb{I}(o_i \leq t)\delta_i + (0 - r_i)^2 \mathbb{I}(o_i > t)].$$

Unlike the C-index, the Brier score does not depend on the relative rank of the subjects. Therefore, we can define the disparity of the Brier score, ΔBrier, as the maximum of the pairwise absolute difference of Brier scores across all possible pairs of groups.

5. Experiments

We performed comprehensive experiments to evaluate FAST on the four proposed disparity metrics using three real-world medical datasets: FLChain, Framingham, and SUPPORT. We followed the `SurvSet` repository (Drysdale, 2022) protocols to preprocess the datasets. Table 1 shows a summary of these datasets; details are in Appendix D.1.

We implemented FAST on three baselines (fairness-unaware) neural survival models (DeepCox, DNNSurv, and DeepHit), all of which have been shown to achieve competitive overall predictive performance in the survival analysis literature. We also compared our method with existing fair survival analysis models: GFDeepCox and IFDeepCox (Keya et al., 2021), GFDNNSurv, and IFDNNSurv (Rahman and Purushotham, 2022), and DRODeepCox (Hu and Chen, 2022). Details for all of the implementations can be found in Appendix D.2 and D.3.

5.1. Comparative Experiments

We split each dataset into three mutually exclusive sets, 60% for training, 20% for validation, and 20% as testing sets. We used the training set to estimate the models and selected hyperparameters using the validation set. We then evaluated the models' performances and the disparities in the testing set. We calculated each metric for time points ranging from the 1st to the 99th percentile and then reported the average of those values. Detailed information, including neural network structures and optimization settings, is in Appendix D.3.

We implemented our hyperparameter selection strategy, which we introduced in Section 3.5, by setting an acceptance margin of $\pm 5\%$. In particular, we selected the hyperparameter that achieves the smallest ΔBrier while ensuring that overall Brier scores do not increase by more than 5%. We choose to use the Brier score as it is a commonly used metric that measures both calibration and concordance (Haider et al., 2020).

In Table 2, all group fairness-encouraging methods, except for individual fairness models, effectively decreased the group fairness metric Gf (defined as the maximum difference of expected per-group survival functions) as expected. We present trade-off curves between the Gf and performance metrics in Figures A5, A6, and A7 to show the effectiveness of each method in decreasing Gf. Moreover, FAST not only improved parity measured in Gf but also effectively decreased disparities in most of the metrics while maintaining the Brier score within the pre-defined acceptance margin. Similar levels of (or even better) efficacy of FAST were observed for the baselines, demonstrating that FAST can be flexibly used for all three types of baseline survival models. In contrast, GFDeepCox, IFDeepCox, and DRODeepCox achieved comparable performance in reducing disparities in FLChain but did not decrease ΔBrier. Moreover, they showed substantially worse overall predictive performance in terms of the Brier score (marked orange when outside the acceptable margin). This may be due to GFDeepCox, IFDeepCox, and DRODeepCox placing emphasis on equity in proportional hazards but not taking the baseline hazards into account. GFDNNSurv and IFDNNSurv worked almost as efficiently as ours in both encouraging predictive parity while not losing their efficacy because they encourage parity of the pseudo-survival probabilities, unlike GFDeepCox, IFDeepCox, and DRODeepCox.

Dataset	Model	Disparity(↓)					Performance		
		GF	ΔBRIER	ΔECE	ΔCI^s	ΔCI^p	BRIER(↓)	ECE(↓)	CI(↑)
FLChain	Vanilla DeepCox	1.49	1.20	1.16	2.03	1.30	8.81	2.69	80.07
	FASTDeepCox (Ours)	-70.8%	-29.5%	22.7%	-18.4%	-10.1%	1.2%	29.2%	-0.4%
	GFDeepCox	-51.3%	-15.0%	104.2%	-12.0%	-19.8%	13.9%	148.0%	-2.9%
	IFDeepCox	-50.9%	-15.2%	118.2%	-12.5%	-19.9%	13.8%	150.6%	-2.9%
	DRODeepCox	-50.7%	-15.2%	106.4%	-11.9%	-20.4%	13.6%	149.8%	-2.9%
	Vanilla DNNSurv	1.20	1.19	1.26	2.61	0.75	8.66	3.55	79.55
	FASTDNNSurv (Ours)	-68.6%	-31.3%	-19.0%	8.4%	16.7%	-0.9%	146.9%	-0.3%
	GFDNNSurv	-50.0%	-8.9%	-16.5%	17.1%	37.8%	-0.4%	4.3%	0.1%
	IFDNNSurv	-2.2%	-0.4%	7.6%	-0.2%	8.8%	0.7%	3.3%	-0.2%
	Vanilla DeepHit	1.12	0.93	1.15	2.86	0.56	8.19	5.83	79.08
	FASTDeepHit (Ours)	-63.5%	-20.2%	39.6%	20.9%	5.6%	5.4%	137.3%	-1.2%
Framingham	Vanilla DeepCox	0.47	4.01	2.19	4.81	1.69	10.38	4.05	68.48
	FASTDeepCox (Ours)	-3.6%	-1.5%	24.1%	-6.2%	0.0%	-0.2%	-2.9%	-0.2%
	GFDeepCox	-25.4%	7.0%	84.5%	-18.8%	36.5%	36.5%	12.7%	-9.1%
	IFDeepCox	-45.0%	3.6%	65.1%	-15.8%	105.1%	13.2%	48.4%	-15.5%
	DRODeepCox	-50.1%	3.8%	84.5%	-12.8%	103.7%	14.1%	58.7%	-15.7%
	Vanilla DNNSurv	0.51	3.40	5.69	4.78	1.19	8.48	16.84	63.93
	FASTDNNSurv (Ours)	-8.7%	-24.7%	25.0%	-37.0%	98.1%	4.6%	19.0%	-2.7%
	GFDNNSurv	-22.8%	0.2%	5.6%	-1.5%	87.3%	0.2%	0.5%	-0.2%
	IFDNNSurv	0.1%	0.0%	-0.3%	-2.7%	-0.1%	0.0%	-0.1%	0.0%
	Vanilla DeepHit	0.67	2.15	7.48	3.01	3.02	8.84	21.32	63.07
	FASTDeepHit (Ours)	-1.4%	-4.1%	4.3%	0.8%	-6.3%	2.0%	4.8%	-0.1%
SUPPORT	Vanilla DeepCox	0.60	3.78	7.99	3.42	2.80	8.99	7.10	84.46
	FASTDeepCox (Ours)	-22.8%	-2.2%	5.4%	1.2%	-2.4%	1.0%	1.4%	-0.3%
	GFDeepCox	-29.3%	7.3%	18.9%	44.0%	86.4%	38.6%	22.2%	-7.4%
	IFDeepCox	-27.5%	9.3%	24.9%	44.0%	86.3%	37.7%	20.9%	-7.4%
	DRODeepCox	-26.5%	6.4%	38.0%	45.1%	86.1%	37.8%	19.3%	-7.4%
	Vanilla DNNSurv	0.55	3.34	8.33	4.65	3.18	7.86	4.73	83.10
	FASTDNNSurv (Ours)	-19.3%	-1.6%	2.5%	-17.3%	-9.4%	1.2%	20.0%	0.5%
	GFDNNSurv	-31.0%	-2.9%	5.3%	-7.8%	2.6%	1.5%	8.6%	0.5%
	IFDNNSurv	-7.7%	-1.3%	1.8%	3.9%	4.2%	-0.4%	10.2%	-0.4%
	Vanilla DeepHit	0.53	2.79	8.29	7.62	3.89	7.42	4.07	82.03
	FASTDeepHit (Ours)	-24.7%	-5.6%	4.3%	3.5%	5.6%	2.0%	36.7%	0.4%

Table 2: Experimental results in performance and disparity metrics from three benchmark datasets. We report actual performances as percentages for the plain baseline methods (Vanilla DeepCox, DNNSurv, and DeepHit). In contrast, for the fairness-aware methods, we show the relative change of each metric compared to its baseline method. Each number represents the average performance across 10 repeated runs. The ↓ symbol indicates that lower values are better, while the ↑ symbol implies that higher values are better. The best results for each baseline method in each dataset are highlighted in Green. Orange highlights (in the performance columns) mean each fair method's performance is outside the 5% margin.

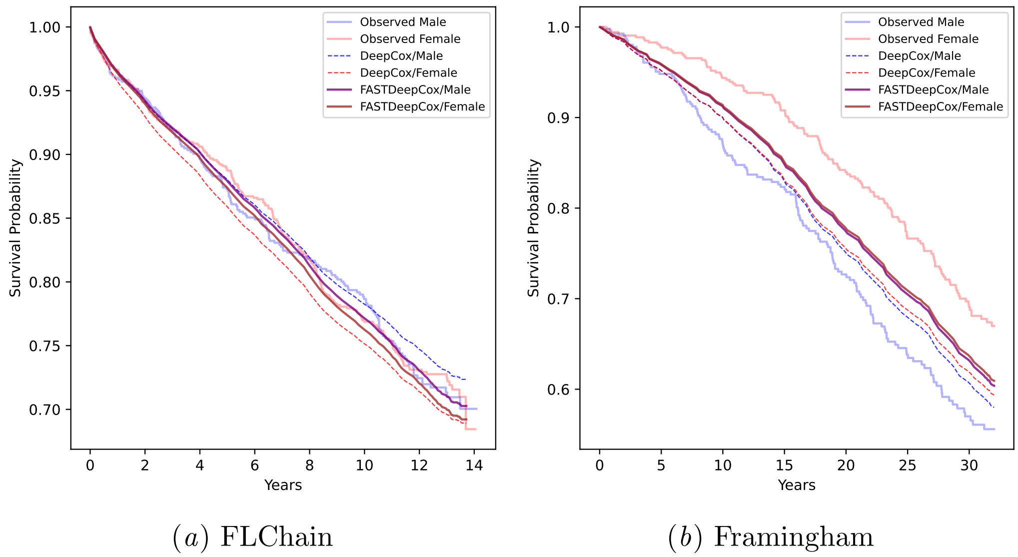
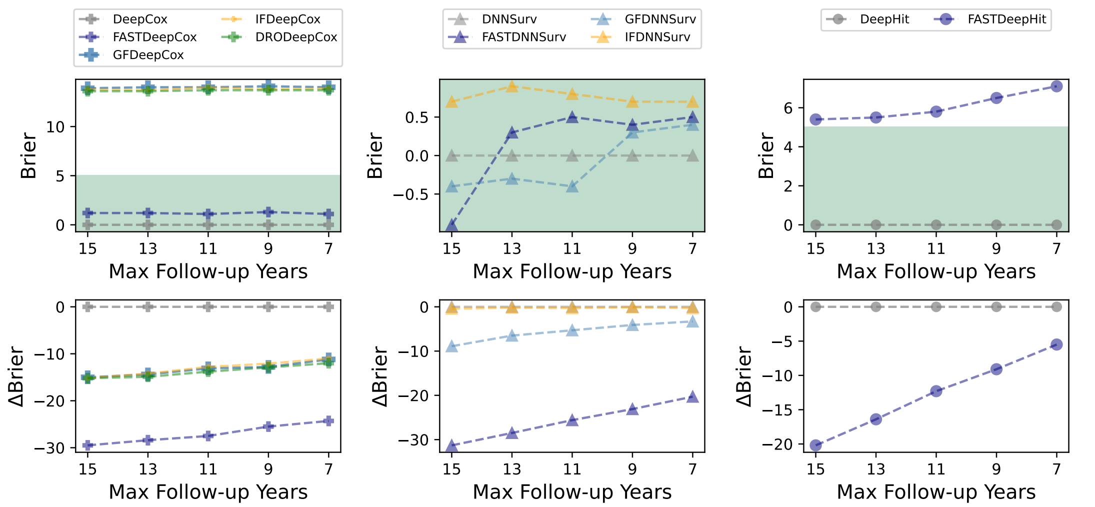


Figure 3: **Improved Survival Predictions from FAST.** The figure compares the observed survival curves (by Kaplan-Meier), predicted survival curves by DeepCox (blue and red dash lines), and predicted survival curves by FAST-DeepCox (purple and brown solid lines) for each sensitive attribute group in the FLChain and Framingham datasets. FAST improves the model's predictions for females and males in FLChain, resulting in predicted survival curves that are closer to their observed counterparts. In Framingham, FAST reduces the bias in the DeepCox prediction, moving the predicted curve towards the middle of the two observed curves. The prediction improvements by FAST from other methods (DNNSurv and DeepHit) are shown in Figures A3 and A4.

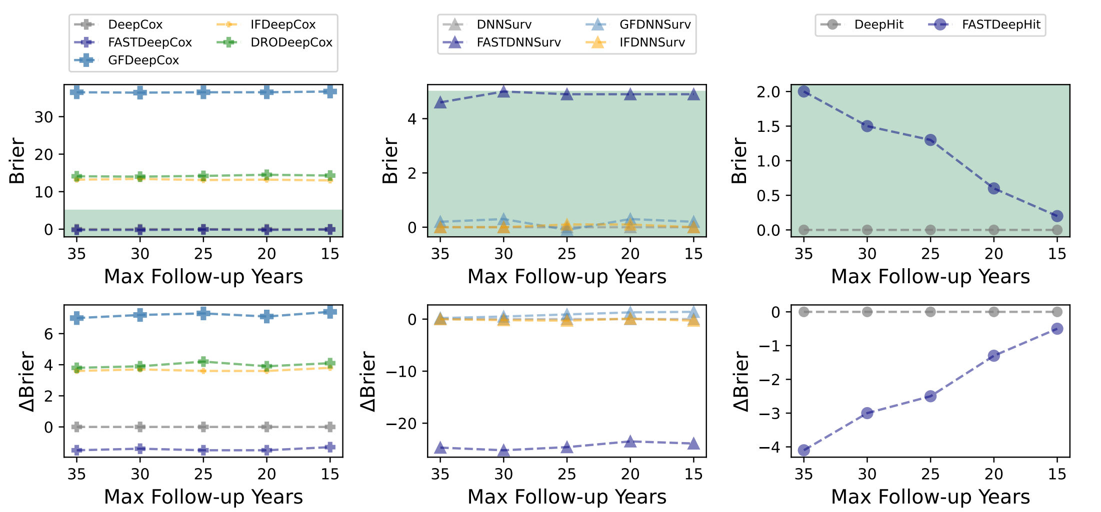
For FAST, γ selected by each metric generally achieved the desired performance for that metric and other metrics in the test set. Although the best performances of a specific metric in the test data are usually achieved by the γ selected by the same metric, γ_{Brier} empirically achieved satisfactory performances on calibration and concordance.

5.2. FAST Robustness against Disparity in Censoring

In survival analysis, a unique but common source of disparity is the difference in proportions of censored subjects between the sensitive attribute groups. If one group has a shorter survival time and/or shorter follow-up time, it will lead to a higher proportion of censoring, which may result in worse prediction performance for that group. To study the robustness of FAST, we designed an experiment on FLChain and Framingham datasets to simulate this scenario. In FLChain, both females and males were followed up to 15 years from the study entry. We truncated the maximum follow-up years for males to $C_{\max} = 15$, 13, 11, 9, and 7 years in the training and validation datasets. In the truncated dataset, we set a male subject i to censored status $(o_i, \delta_i) = (C_{\max}, 0)$ if $o_i > C_{\max}$. Thus, the proportion of censoring in males increased and became more disparate from females with decreasing $C_{\max}$.



(*a*) FLChain



(*b*) Framingham

Figure 4: Model performances and disparities (relative changes in percentage compared to the corresponding baselines) versus max follow-up years for males for the FLChain and Framingham datasets averaged over 5 test sets. The green shades represent the 5% acceptance margin of the Brier score.

We present the follow-up distribution for the event and censored individuals in Figure A11. We implemented the various methods on the truncated dataset and selected the optimal γ_{Brier} by setting a 5% acceptance margin. In Framingham, both females and males were followed up to 35 years from the study entry. We truncated the maximum follow-up years for males (minority group) to $C_{\max} = 35$, 30, 25, 20, and 15 years in the training and validation datasets, as same as the previous. Follow-up time distributions are presented in Figure A12.

Figure 4 confirms our hypothesis, in both datasets, that disparate follow-up time directly results in disparate Brier scores from the fairness-unaware survival models. However, FAST methods achieve consistent and robust improvements in Brier score disparities across the range of truncated follow-up times while maintaining acceptable overall Brier scores (marked green margins). In contrast, the fairness-aware competitors are not as effective in decreasing disparities or maintaining overall Brier scores. We also noticed that the disparate follow-up time has relatively less impact on concordance metrics (Figures A13 and A15 in Appendix) but worsens the calibration (Figures A14 and A16 in Appendix). We conjecture that this is because disparate follow-up time impacts disparity in baseline survival functions and hazard ratio estimates. Therefore, fairness approaches that impose parity on the hazard parameters (such as GFDeepCox, IFDeepCox, and DRODeepCox) are less effective in this setting. This empirically demonstrates the advantage of directly achieving parity on the predicted survival time via a nonparametric MI estimation, which makes FAST robust against a range of possible disparities, including disparities in baseline survival probabilities and hazard functions.

6. Conclusion and Discussion

We developed a Fair Survival Time Prediction (FAST) method, via a mutual information penalty term to learn survival time predictions, which is compatible with survival analysis formulations with differentiable log-likelihood functions, including recent deep survival analysis models. We also proposed a series of metrics to evaluate parity in survival time predictions. FAST shows empirical consistency in improving prediction parity while maintaining overall prediction performance. Furthermore, we demonstrate that FAST is robust to the presence of disparity in censoring and follow-up times, which is common and essential for survival analysis applications.

Future Directions An interesting extension of work would be to consider other fairness notions such as *Equalized Odds* (Hardt et al., 2016). This will relax the assumption requiring independence between true survival time and A, which has the implications of considering diseases that have different impacts on different gender/races due to genetic reasons. However, such extensions are not trivial as the true survival time T is not observed because of C. Another potential future direction is to investigate the impact of the informative censoring (Lagakos, 1979) on the disparity in survival analysis and to develop a method to mitigate it.

Limitations As our method relies on MINE (Belghazi et al., 2018), which utilizes the deep neural network to estimate mutual information, it cannot be as simplified as penalized linear models. Even if our approach can be combined with a linear survival model, it still requires a nonlinear neural network to estimate MI. Another limitation of our framework is that we did not impose fairness on an individual level, so some individuals might get disadvantaged by considering the fairness between groups. We encourage future work in this field to take this into account.

Societal Impact In areas such as healthcare, ensuring fairness in predictive models is essential given that these models are utilized to make highly sensitive, life-changing, and even life-saving decisions. Most of these models make decisions based on predicted survival time. Our methodology and proposed parity metrics provide a new perspective and framework to systematically improve fairness for survival predictions.

Acknowledgments

We thank the MLHC reviewers for their suggestions on improving the original version of this paper. This work was supported in part by the National Institutes of Health under awards NIH R01-LM013344, R01-AG054467, R01-AG065330, R01-AG065330-02S1, by NIH Agreement No. 1OT2OD032581-01, by NYU Center for the Study of Asian American Health under the NIH/NIMHD grant award #U54MD000538 (HD), by the National Science Foundation under award 1900644 (PS), and by the HPI Research Center in Machine Learning and Data Science at UC Irvine (YC, PS).

References

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 120–129. PMLR, 09–15 Jun 2019.

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*, 39 (21):2714–2742, 2020.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540. PMLR, 10–15 Jul 2018.

HR Byers, HE Landsberg, H Wexler, B Haurwitz, AF Spilhaus, HC Willett, HG Houghton, Glenn W Brier, and Roger A Allen. Verification of weather forecasts. *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, pages 841–848, 1951.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, 2021. doi: 10.1146/annurev-biodatasci-092820-114757.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. In *Advances in Neural Information Processing Systems*, volume 33, pages 19137–19148. Curran Associates, Inc., 2020.

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

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1436–1445. PMLR, 09–15 Jun 2019.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. In *Advances in Neural Information Processing Systems*, volume 34, pages 26740–26753. Curran Associates, Inc., 2021.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Hyungrok Do, Preston Putzel, Axel S Martin, Padhraic Smyth, and Judy Zhong. Fair generalized linear models with a convex penalty. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 5286–5308. PMLR, 17–23 Jul 2022.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Erik Drysdale. SurvSet: An open-source time-to-event dataset repository. *arXiv preprint arXiv:2203.03094*, 2022.

Stephanie S. Gervasi, Irene Y. Chen, Aaron Smith-McLallen, David Sontag, Ziad Obermeyer, Michael Vennera, and Ravi Chawla. The potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41(2):212–218, 2022.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, Nov 2018.

Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, and Rajesh Ranganath. X-cal: Explicit calibration for survival analysis. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pages 2262–2268, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR, 06–11 Aug 2017.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *The Journal of Machine Learning Research*, 21 (1):3289–3351, 2020.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Frank E Harrell, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, 1984.

Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018.

Shu Hu and George H. Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Proceedings of the 2nd Machine Learning for Health Symposium*, volume 193, pages 62–87. PMLR, 28 Nov 2022.

Shinya Iwase, Taka-aki Nakada, Tadanaga Shimada, Takehiko Oami, Takashi Shimazui, Nozomi Takahashi, Jun Yamabe, Yasuo Yamao, and Eiryo Kawakami. Prediction algorithm for icu mortality and length of stay using machine learning. *Scientific Reports*, 12(1):1–9, 2022.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *The 2nd International Conference on Computer, Control, and Communication*, pages 1–6, 2009.

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): 1–12, 2018.

Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James Foulds. Equitable allocation of healthcare resources with fair survival models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 190–198. SIAM, 2021.

Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 11074–11101. PMLR, 17–23 Jul 2022.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

Matthäus Kleindessner, Samira Samadi, Muhammad Bilal Zafar, Krishnaram Kenthapadi, and Chris Russell. Pairwise fairness for ordinal regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 3381–3417. PMLR, 28–30 Mar 2022.

William A Knaus, Frank E Harrell, Joanne Lynn, Lee Goldman, Russell S Phillips, Alfred F Connors, Neal V Dawson, William J Fulkerson, Robert M Califf, Norman Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995.

Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Dirk R Larson, Matthew F Plevak, Janice R Offord, Angela Dispenzieri, Jerry A Katzmann, and L Joseph Melton III. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 354(13):1362–1369, 2006.

Stephen W Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979.

Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 12917–12930. PMLR, 17–23 Jul 2022.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3384–3393. PMLR, 10–15 Jul 2018.

Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921):999–1008, 2014.

Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, 2021.

Allan H Murphy. Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology*, pages 273–282, 1972.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep cox mixtures for survival regression. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149, pages 674–708. PMLR, 06–07 Aug 2021.

Johan Nilsson, Mattias Ohlsson, Peter Höglund, Björn Ekmehag, Bansi Koul, and Bodil Andersson. The international heart transplant survival algorithm (ihtsa): a new model to improve organ sharing and survival. *PloS One*, 10(3):e0118644, 2015.

Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3(1):1–8, 2020.

Md Mahmudur Rahman and Sanjay Purushotham. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1452–1462, 2022.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2586–2594, 2019.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 232–243. World Scientific, 2020.

Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19598–19615. PMLR, 17–23 Jul 2022.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 20156–20175. PMLR, 17–23 Jul 2022.

Alexander Soen, Ibrahim M Alabdulmohsin, Sanmi Koyejo, Yishay Mansour, Nyalleng Moorosi, Richard Nock, Ke Sun, and Lexing Xie. Fair wrapping for black-box predictions. In *Advances in Neural Information Processing Systems*, volume 35, pages 21615–21627. Curran Associates, Inc., 2022.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2164–2173. PMLR, 16–18 Apr 2019.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized score transformation for fair classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1673–1683. PMLR, 26–28 Aug 2020.

Zhenhuan Yang, Yan Lok Ko, Kush R. Varshney, and Yiming Ying. Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11909–11917, Jun. 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970. PMLR, 20–22 Apr 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 17–19 Jun 2013.

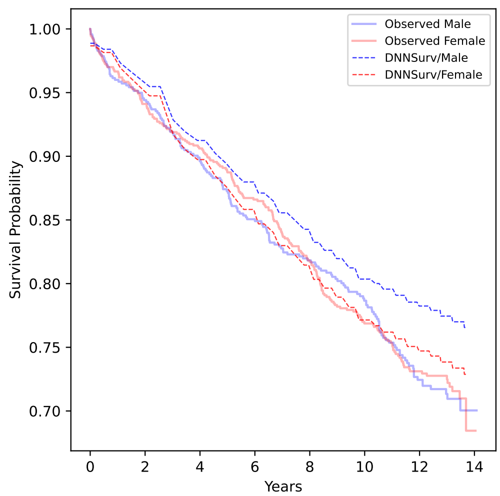
Wenbin Zhang and Jeremy C. Weiss. Fair decision-making under uncertainty. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 886–895, 2021.

Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12235–12243, 2022.

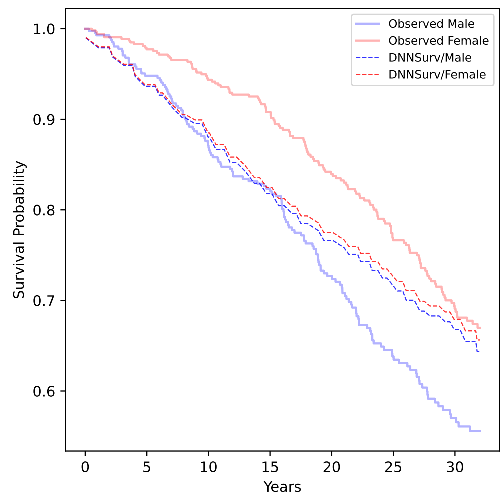
Tianhang Zheng and Baochun Li. Infocensor: An information-theoretic framework against sensitive attribute inference and demographic disparity. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, page 437–451. Association for Computing Machinery, 2022.

Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15002–15012, October 2021.

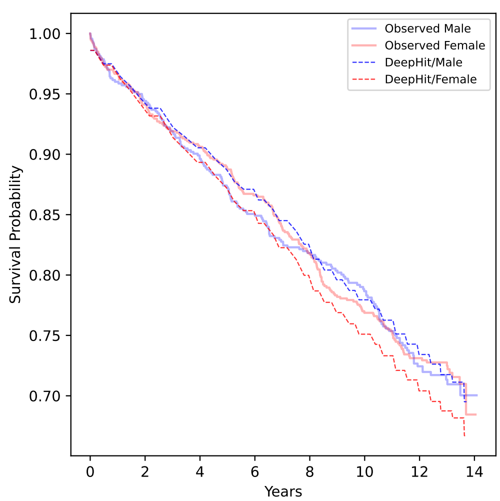
Appendix A. Comparison of the Observed and Predicted Survival Curves



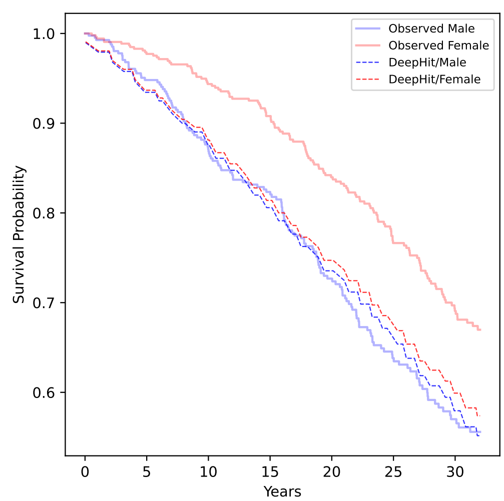
(a) FLChain-DNNSurv



(b) Framingham-DNNSurv



(c) FLChain-DeepHit



(d) Framingham-DeepHit

Figure A1: This figure compares observed survival curves (generated using Kaplan-Meier analysis) with predicted curves (generated using DNNSurv and DeepHit models) for two datasets: FLChain and Framingham. The results indicate that for FLChain, both DNNSurv and DeepHit models tend to underestimate the survival probability (or overestimate the risk of the event) for females. Similarly, for Framingham, both models tend to overestimate the survival probability for females. These findings suggest that the prediction disparity is not specific to the Cox proportional hazard model but rather a systematic issue that is inherent to the data.

Appendix B. Further Discussions

B.1. Alternative Statistical Dependence Metrics (Supplement for Section 3.1)

Here, we give a brief discussion on metrics to quantify statistical dependence other than mutual information. DP requires the predicted survival time $\mathcal{T}_\theta$ to be statistically independent of the sensitive attributes A. Given the set of possible sensitive attributes $\mathcal{A} = \{\alpha_1, \cdots, \alpha_K\}$, the independence of $\mathcal{T}_\theta$ and A translates into $\mathcal{T}_\theta = (\mathcal{T}_\theta | A = \alpha_k)$ for all $k = 1, \cdots, K$. Therefore, $\mathcal{T}_\theta$ is independent of A if and only if $\mathbb{D}(\mathcal{T}_\theta, \mathcal{T}_\theta | A = \alpha_k) = 0$ for all $k = 1, \cdots, K$, where $\mathbb{D}$ is a distance or divergence between two probability distributions, such as Kullback-Leibler divergence, Jensen-Shannon divergence, total variation distance, Wasserstein distance, and maximum mean discrepancy (Gretton et al., 2012). Therefore, we can include one of the following terms as a DP encouraging penalty:

$$\sum_{k=1}^{K} \hat{\mathbb{D}}(\mathcal{T}_\theta, \mathcal{T}_\theta | A = \alpha_k), \qquad \text{or} \qquad \max_k \hat{\mathbb{D}}(\mathcal{T}_\theta, \mathcal{T}_\theta | A = \alpha_k). \tag{A1}$$

The penalty term equals zero means $\mathcal{T}_\theta$ and A are statistically independent, the same as when mutual information is zero. Therefore, the distances and divergences introduced above can be used as an alternative to achieve DP. However, even if minimizing them toward zero implies the statistical independence of $\mathcal{T}_\theta$ and A, each divergence or distance may generate a different trajectory of solutions obtained by changing the hyperparameter that controls the trade-off between the negative log-likelihood and the penalty term. We leave this as a future direction.

B.2. Comparison Against Fair Representation Learning

In the fairness literature, a line of work encourages the model output to satisfy some fairness criteria by learning a fair representation that can be transferred to fair performances for various downstream tasks. Zemel et al. (2013) proposed the first FRL method to achieve DP for binary classification. Further extensive work has been done recently for sophisticated fairness criteria other than DP, as well as for various downstream tasks (Madras et al., 2018; Roy and Boddeti, 2019; Gupta et al., 2021; Kim et al., 2022; Shui et al., 2022). The fair representation learning to achieve the DP aims to learn a function $\psi : \mathcal{X} \to \mathcal{Z}$ such that $\mathbf{Z} \perp\!\!\!\perp A$, and build a model $\phi : \mathcal{Z} \to \mathcal{Y}$ for the downstream task (usually prediction model) on top of ψ. This approach may look similar to our approach as encouraging the independence of $\mathbf{Z} = \psi(\mathbf{X})$ and A transfers to encouraging the independence of $(\phi \circ \psi)(\mathbf{X}) = \hat{Y}$ and A.

However, in the survival analysis setting, the predicted survival time $\mathcal{T}_\theta$ is not exactly the outcome of a model (or a network). In general, the model output is the hazard function, event probability, or survival probability. Therefore, the model outcome is independent of the sensitive attribute is totally different from the predicted survival time is independent of the sensitive attribute. Moreover, some survival models (e.g., Cox proportional hazard model or accelerated failure time model) have the baseline hazard function, which is totally separated from the model output. For such models, encouraging any kind of fairness, including independence, of the model output will not result in the fairness of the predicted survival time because it does not account for the baseline hazard.

B.3. Comparison Against Group Fair DeepCox

Group Fair DeepCox (Keya et al., 2021) encourages the expected proportional hazard to be the same across the sensitive groups, that is,

$$\mathbb{E}[\exp f_\theta(\mathbf{X}|A=\alpha_k)] = \mathbb{E}[\exp f_\theta(\mathbf{X})], \tag{A2}$$

for all $k = 1, \cdots, K$. We note that this criterion is different from DP, which requires the predicted survival time to be independent of the sensitive attribute. Equalizing the expectation of the group-conditional hazard functions across all groups does not necessarily result in the same distributions of group-conditional predicted survival times (but the reverse is true). Furthermore, (A2) is the equivalence in *proportional hazard* without considering the baseline hazard, so the criterion does not imply the expected hazard functions being the same. Thus, (A2) is a much weaker condition than DP, and it cannot account for the disparity involving the baseline hazard function.

Appendix C. FAST Formulations (Supplement for Sections 3.3 and 3.4)

We introduce complete formulations of our FAST approaches by specifying the log-likelihood functions for some representative survival analysis models.

C.1. Notations

Dataset $\{(\mathbf{x}_i, a_i, o_i, \delta_i) : i = 1, \cdots, n\}$, where $\mathbf{x}_i$ is the covariate vector, a_i is the sensitive attribute, o_i is the observed time (minimum of time to event or censoring), and δ_i is the event indicator which takes 1 if the event precedes the censoring and 0 otherwise.

Model $f_\theta : \mathcal{X} \to \mathbb{R}$, parameterized by θ. Either linear model or neural network.

Log-likelihood

$$\ell(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i \log \lambda_\theta(o_i|\mathbf{x}_i) - \Lambda_\theta(o_i|\mathbf{x}_i)\Big]. \tag{A3}$$

In the case of discrete-time models (such as DeepHit), the negative log-likelihood becomes

$$\ell(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i \log P(\mathcal{T}_\theta(\mathbf{x}_i) = o_i) - (1-\delta_i) P(\mathcal{T}_\theta(\mathbf{x}_i) > o_i)\Big]. \tag{A4}$$

C.2. Cox Proportional Hazard Model (CPH)

Log-likelihood Cox proportional hazard model (Cox, 1972) is based on the proportional hazard assumption, which is $\lambda_\theta(t|\mathbf{x}) = \lambda_0(t)\lambda_\theta(\mathbf{x})$, where λ_0 is the baseline function and λ_θ is the proportional hazard function. We introduce a neural network $f_\theta : \mathbf{x} \mapsto \log \lambda_\theta(\mathbf{x})$ which outputs a log-hazard value given an input $\mathbf{x}$.

The most widely used loss function to learn Cox proportional hazard models, for both linear (Cox, 1972) and deep neural network (Katzman et al., 2018) models, is the negative partial log-likelihood, which is defined as follows:

$$-\ell_p(\theta; \mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\Big[f_\theta(\mathbf{x}_i) - \log\Big(\sum_{j\in\mathcal{R}_i}\exp f_\theta(\mathbf{x}_j)\Big)\Big], \tag{A5}$$

where $\mathcal{R}_i = \{j : \delta_j = 1, o_i < o_j\}$ is the at-risk set at the individual i event time o_i. However, the stochastic gradient approach for negative partial log-likelihood leads to a biased estimate of the full batch gradient. Hence, instead, we take the approach that first estimates the baseline hazard function, then plugs it into the full log-likelihood, and minimizes the negative log-likelihood. Let $\hat{\lambda}_0$ and $\hat{\Lambda}_0$ be the *estimated* baseline hazard and cumulative baseline hazard functions, respectively, which are estimated solely with $\{(o_i, \delta_i) : i = 1, \cdots, n\}$. Then, we have the following negative log-likelihood minimization problem:

$$\min_\theta -\frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i\big(\log\hat{\lambda}_0(o_i) + f_\theta(\mathbf{x}_i)\big) - \hat{\Lambda}_0(o_i)\exp f_\theta(\mathbf{x}_i)\Big], \tag{A6}$$

whose stochastic gradient with respect to θ is unbiased. The quality of the full log-likelihood depends on the choice and estimation of baseline hazard functions, but we found that the results are fairly consistent with respect to the choice of baseline hazard function using neural networks with sufficiently large capacity. In this paper, we used the Weibull baseline hazard function, however, any popular baseline hazard function, including nonparametric ones such as Kaplan-Meier (Kaplan and Meier, 1958) can be used as well.

Sampling From $\mathcal{T}_\theta(\mathbf{x})$ In the case of the Cox PH model, the inverse CDF has a closed form if we use a parametric baseline hazard function. Thus, we use inverse transformation sampling. As CDF of $\mathcal{T}_\theta(\mathbf{x})$ is given as $F_\theta(t|\mathbf{x}) = P(\mathcal{T}_\theta(\mathbf{x}) \le t) = 1 - \exp(-\hat{\Lambda}_0(t)\exp f_\theta(\mathbf{x}))$, we can easily find the inverse CDF. If we use the exponential baseline hazard function $\lambda_0(t) = \frac{1}{\lambda}$, then $\mathcal{T}_\theta(\mathbf{x}) \sim \texttt{Exponential}(\lambda/\exp f_\theta(\mathbf{x}))$,

$$F_\theta(F_\theta^{-1}(u|\mathbf{x})|\mathbf{x}) = 1 - \exp(-\hat{\Lambda}_0(t)\exp f_\theta(\mathbf{x})) = 1 - \exp\left(-\frac{t}{\hat{\lambda}}\exp f_\theta(\mathbf{x})\right),$$

and therefore,

$$F_\theta^{-1}(u|\mathbf{x}) = -\frac{\hat{\lambda}\log(1-u)}{\exp f_\theta(\mathbf{x})} = -\hat{\lambda}\log(1-u)\exp(-f_\theta(\mathbf{x})).$$

To get sample τ_i drawn from $\mathcal{T}_\theta(\mathbf{x}_i)$, we first draw $u_i \sim \texttt{Uniform}(0,1)$, and compute $F_\theta^{-1}(u_i|\mathbf{x}_i)$.

If we use the Weibull baseline hazard function $\lambda_0(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1}$, $\mathcal{T}_\theta(\mathbf{x}) \sim \texttt{Webuill}\left(\lambda\exp(-\frac{1}{\rho}f_\theta(\mathbf{x})), \rho\right)$. Even for nonparameteric baseline hazard functions, for example, the Kaplan-Meier estimator, we can easily obtain the inverse function from the step function that defines the Kaplan-Meier estimator, as described in A2.

C.3. DeepHit

Log-likelihood DeepHit (Lee et al., 2018) directly models the event probability at each discretized time using a neural network $f_\theta : \mathbf{x} \mapsto P(\mathcal{T}_\theta(\mathbf{x}) = t)$. This leads to the following negative log-likelihood minimization problem:

$$\min_\theta -\frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i \log P(\mathcal{T}_\theta(\mathbf{x}_i) = o_i) - (1-\delta_i)P(\mathcal{T}_\theta(\mathbf{x}_i) > o_i)\Big]. \tag{A7}$$

Sampling From $\mathcal{T}_\theta(\mathbf{x})$ DeepHit defines event probability by a categorical distribution, so we can easily draw the time interval of an event from the distribution. However, sampling from a discrete probability distribution is not a differentiable operation. Thus, we used the Gumbel-Softmax trick (Huijben et al., 2022) in our implementation that allows differentiable sampling operation.

C.4. DNNSurv

Log-likelihood DNNSurv (Zhao and Feng, 2020) is a pseudo value (Andersen and Pohar Perme, 2010)-based deep survival model. It first estimates *pseudo survival function* for each individual using the Jackknife method with the Kaplan-Meier estimator. It is known that such pseudo value estimators are asymptotically unbiased for both censored and uncensored observations. Subsequently, they build a neural network model $f_\theta : \mathbf{x} \mapsto P(\mathcal{T}_\theta(\mathbf{x}) > t) \in [0, 1]$ to predict the pseudo survival functions correctly. In Zhao and Feng (2020), the mean square error (MSE) is used as the loss function, but the MSE for pseudo value-based survival analysis can have convergence issues in heavily censored settings (Rahman and Purushotham, 2022). We used the modified loss function in Rahman and Purushotham (2022) and defined the problem as

$$\min_\theta -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\Big[\widehat{S}_i(t_j)(1 - 2P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)) + P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)^2\Big], \tag{A8}$$

where $\{t_1, \cdots, t_J\}$ is a set of discrete time points that we define before performing Jackknife.

Sampling From $\mathcal{T}_\theta(\mathbf{x})$ Unlike CoxPH and DeepHit, sampling predicted survival times is not straightforward. Thus, we propose a differentiable sampling from the empirical cumulative distribution function. Given an individual, we first evaluate the empirical CDF on a grid of time points. The grid should cover the range from 0 to a sufficiently large value (e.g., the max follow-up time of the dataset). Then, we draw $u \sim \texttt{Uniform}(0, 1)$ as in the inverse CDF sampling, find the closest empirical CDF value, then find the corresponding input value. We illustrate this procedure in Figure A2. To make the operation *differentiable*, we use the Gumbel-Softmin trick when we find the closest empirical CDF value. This technique not only applies to FAST-DNNSurv, but can be applied to any of the existing survival models.

C.5. Cox-Time

Log-likelihood We also consider the Cox-time model that assumes the relative risk is not only a function of covariates $\mathbf{x}$ but also a function of t, that is, $\lambda_\theta(t|\mathbf{x}) = \lambda_0(t)\exp f_\theta(t, \mathbf{x})$. Thus, for the Cox-time model, a neural network takes both covariate vector $\mathbf{x}_i$ and time t as inputs and outputs the log-proportional hazard, that is, $f_\theta : (\mathbf{x}, t) \mapsto \log \lambda_\theta(t, \mathbf{x})$. Again, as in CoxPH model, we first estimate the baseline hazard function and plug it into the full log-likelihood (so that it is different from that of Kvamme et al. (2019)), leading to the following optimization problem:

$$\min_\theta -\frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i\big(\log \lambda_0(o_i) + f_\theta(o_i, \mathbf{x}_i)\big) - \Lambda_0(o_i)\exp f_\theta(o_i, \mathbf{x}_i)\Big]. \tag{A9}$$

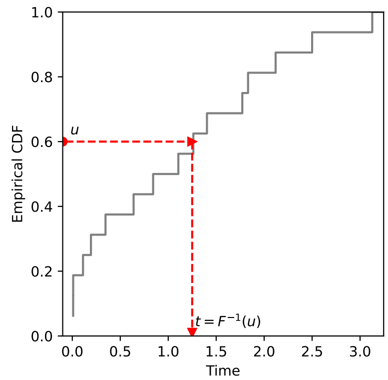


Figure A2: Sampling From Empirical Cumulative Distribution Function. We (1) sample $u \sim \texttt{Uniform}(0, 1)$; (2) find the closest empirical CDF value; (3) find the corresponding t. In step (2), we substitute the Gumbel-Softmin trick for `argmin` to make the operation differentiable.

Unlike the CoxPH model, Cox-time does not allow closed forms for relative hazard or cumulative relative hazard functions. Thus, we use the Riemann sum to approximate the cumulative hazard function as follows:

$$\Lambda_\theta(o_i|\mathbf{x}_i) = \int_0^{o_i} \lambda_\theta(t|\mathbf{x}_i)dt = \int_0^{o_i} \lambda_0(t) \exp f_\theta(t, \mathbf{x}_i)dt \approx \sum_{j=1}^{m} \lambda_0(s_j) \exp f_\theta(s_j, \mathbf{x}_i)(s_j - s_{j-1}), \tag{A10}$$

where $\{s_j : 0, \cdots, m\}$ is a properly defined grid such that $s_0 = 0$ and $s_m = o_i$.

Sampling From $\mathcal{T}_\theta(\mathbf{x})$ No closed form CDF is available for Cox-Time. We can do this as same as in the DNNSurv case, as described in Figure A2.

C.6. Accelerated Failure Time Model

Log-likelihood We use the approach of Zhong et al. (2021) for the deep accelerated failure time model, which assumes $\lambda_\theta(t|\mathbf{x}) = \lambda_0(t \exp f_\theta(\mathbf{x})) \exp f_\theta(\mathbf{x})$. The minimization of pseudo-likelihood of Zhong et al. (2021) is defined as:

$$\begin{aligned} \min_\theta -\frac{1}{n}\sum_{i=1}^{n} \delta_i \Bigg[&-\log o_i + \log\Big[\frac{1}{n\nu}\sum_{j=1}^{n} \delta_j \phi\Big(\frac{\log o_i + f_\theta(\mathbf{x}_i) - \log o_j - f_\theta(\mathbf{x}_j)}{\nu}\Big)\Big] \\ &- \log\Big[\frac{1}{n}\sum_{j=1}^{n} \Phi\Big(\frac{\log o_i + f_\theta(\mathbf{x}_i) - \log o_j - f_\theta(\mathbf{x}_j)}{\nu}\Big)\Big]\Bigg], \end{aligned} \tag{A11}$$

where ϕ and Φ are density function and cumulative distribution function of the standard normal distribution, and ν is the bandwidth of the Gaussian kernel (typical choice is

$\nu = 1.3n^{-0.2}$). Moreover, its baseline hazard function is given as

$$\hat{\lambda}_0(t) = \frac{\frac{1}{t}\sum_{i=1}^{n} \delta_i \frac{1}{t}\phi(\frac{\log o_i + f_\theta(\mathbf{x}_i) - \log t}{\nu})}{\sum_{i=1}^{n} \int_{-\infty}^{\log o_i + f_\theta(\mathbf{x}_i) - \log t} \frac{1}{\nu}\phi(\frac{s}{\nu})ds}, \tag{A12}$$

and the cumulative baseline hazard function is given as

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(s)ds. \tag{A13}$$

Finally, the survival function is given as

$$\hat{S}(t|\mathbf{x}) = \exp\big(-\hat{\Lambda}_0(t \exp f_\theta(\mathbf{x}_i)\big). \tag{A14}$$

Sampling From $\mathcal{T}_\theta(\mathbf{x})$ We cannot find a closed-form CDF for the deep AFT. We can do this as same as in the DNNSurv case, as described in Figure A2.

We further note that our approach can be applied to deep extended hazard model (Zhong et al., 2021), piecewise constant hazard model (Kvamme and Borgan, 2021), and SODEN (Tang et al., 2022).

Appendix D. Experiments (Supplement for Section 5)

D.1. Datasets

For all datasets, as presented in the main text, we followed the protocol of `SurvSet` repository (Drysdale, 2022). We imputed the missing values to the median and mode values for numerical and categorical variables, respectively. We then performed one-hot encoding for categorical variables. All covariates taking continuous values were standardized to have zero-mean and unit variance.

FLChain The free light chain (FLChain) dataset resulted from a study about the relationship between serum FLC and mortality. Of the 7,874 patients, 2,169 patients (27.5%) died, and the remaining 5,705 patients (72.5%) were censored. It includes covariates such as age, serum creatinine, the presence of monoclonal gammopathy, etc. For this dataset, gender is the sensitive attribute.

SUPPORT The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) dataset was obtained from a study to understand survival over 180 days for seriously ill hospitalized patients. Of the 9,105 patients, 6,201 patients (68.1%) patients died, with a median survival time of 58 days, and the other 2,904 patients (31.9%) were censored. It contains covariates, including age, gender, education, income, physiological measurements, and co-morbidity information, etc. For this dataset, race (White, Black, Hispanic, and Asian/Other) is the sensitive attribute.

Framingham The Framingham dataset collected from Framingham Heart Study (FHS) was performed to characterize major risk factors that contribute to cardiovascular disease. Of the 4,699 patients, 1,473 patients (31.3%) died, and the remaining 3,226 patients (68.7%) were censored. The dataset includes covariates such as systolic and diastolic blood pressure, age, body mass index (BMI), etc., where gender is the sensitive attribute.

D.2. Competitive Methods (Supplement for Section 5)

GFDeepCox GFDeepCox stands for Group Fairness DeepCox, proposed by Keya et al. (2021), which aims to achieve group fairness for the CoxPH model. It encourages the average proportional hazard of each sensitive group to be similar. The problem is defined as:

$$\min_{\theta} - \sum_{i=1}^{n} \delta_i \left[f_\theta(\mathbf{x}_i) - \log \Big(\sum_{j \in \mathcal{R}(o_i)} \exp f_\theta(\mathbf{x}_j) \Big) \right] + \gamma \max_k \Big| \mathbb{E}[\exp f_\theta(\mathbf{X}_k)] - \mathbb{E}[\exp f_\theta(\mathbf{X})] \Big|, \tag{A15}$$

where $\mathcal{R}(t)$ is the at-risk set at time t.

IFDeepCox IFDeepCox stands for Individual Fairness DeepCox, proposed by Keya et al. (2021), which aims to achieve individual fairness for the CoxPH model. It encourages the proportional hazard to be similar for individuals who have similar inputs. The problem is defined as:

$$\min_{\theta} - \sum_{i=1}^{n} \delta_i \left[f_\theta(\mathbf{x}_i) - \log \Big(\sum_{j \in \mathcal{R}(o_i)} \exp f_\theta(\mathbf{x}_j) \Big) \right] + \gamma \sum_{i,j} \max\{0, |\exp f_\theta(\mathbf{x}_i) - \exp f_\theta(\mathbf{x}_j)| - D(\mathbf{x}_i, \mathbf{x}_j)\}, \tag{A16}$$

where $D(\mathbf{x}_i, \mathbf{x}_j)$ is a distance function.

For GFDeepCox and IFDeepCox, we worked on top of the author's implementation (`https://github.com/kkeya1/FairSurv/`).

GFDNNSurv GFDNNSurv stands for Group Fairness DNNSurv, proposed by Rahman and Purushotham (2022), which aims to achieve group fairness for the DNNSurv (deep pseudo value-based survival model). In the original paper, the model is referred to as FGDP. The formulation to learn the model is:

$$\begin{aligned} \min_{\theta} \; - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \Big[\widehat{S}_i(t_j)(1 - 2P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)) + P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)^2 \Big] \\ + \gamma \sum_{j=1}^{J} \Big| \mathbb{E}[P(\mathcal{T}_\theta(\mathbf{X}_k) > t_j)] - \mathbb{E}[P(\mathcal{T}_\theta(\mathbf{X}) > t_j)] \Big|, \end{aligned} \tag{A17}$$

where $\{t_1, \cdots, t_J\}$ is a set of discrete time points that we define before performing Jackknife.

IFDNNSurv GFDNNSurv stands for Individual Fairness DNNSurv, proposed by Rahman and Purushotham (2022), which aims to achieve individual fairness for the DNNSurv (deep pseudo value-based survival model). In the original paper, the model is referred to as FIDP. The formulation to learn the model is:

$$\begin{aligned} \min_{\theta} \; - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \Big[\widehat{S}_i(t_j)(1 - 2P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)) + P(\mathcal{T}_\theta(\mathbf{x}_i) > t_j)^2 \Big] \\ + \gamma \sum_{l=1}^{L} \sum_{i,j} \max\{0, |P(\mathcal{T}_\theta(\mathbf{x}_i) > t_l) - P(\mathcal{T}_\theta(\mathbf{x}_j) > t_l)| - \alpha D(\mathbf{x}_i, \mathbf{x}_j)\}, \end{aligned} \tag{A18}$$

where $\{t_1, \cdots, t_L\}$ is a set of discrete time points that we define before performing Jackknife, and $D(\mathbf{x}_i, \mathbf{x}_j)$ is a distance function.

Since there is no publicly available code for GFDNNSurv and IFDNNSurv, we implemented the model in our own way.

DRODeepCox DRO for deep Cox proportional hazard model has been proposed by Hu and Chen (2022). We followed the authors' implementation given in `https://github.com/discovershu/DRO_COX`.

D.3. Network Structures and Optimization Details (Supplement for Section 5.1)

Survival Model Network For all three datasets and for all models, we use multilayer perceptrons that have the structure: `Input` – `Dense(64)` – `Dense(64)` – `Dense(64)` – `Output`, where input differs by the dataset and output depends on the choice of the survival model. For instance, for the case of DeepCox `Output` $= 1$, while for DeepHit and DNNSurv, `Output` is the number of discretized time bins. Each dense layer is followed by batch normalization (Ioffe and Szegedy, 2015) and the ReLU activation function (Nair and Hinton, 2010), except for the output layer. For DeepHit and DNNSurv, we discretized the follow-up time into 50 bins with equal percentiles of the dataset.

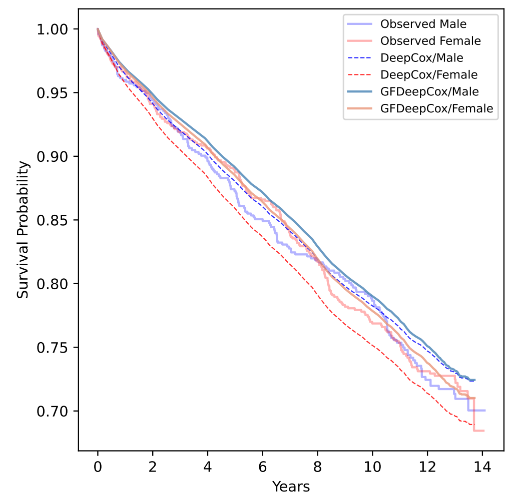
FAST Methods For our FAST approach, we need an additional network to estimate the mutual information of $\mathcal{T}_\theta$ and A. For all datasets, we used multilayer perceptrons that consist of `Input` – `Dense(32)` – `Dense(32)` – `Dense(32)` – `Dense(1)`, where the input has the shape of $K + 1$ because we used one-hot encoded sensitive attributes as input (where K is the number of sensitive groups). Each dense layer is followed by an ELU activation.

Optimization We use Adam (Kingma and Ba, 2014) with weight decay (L2 penalty) of 0.0001 for all methods and datasets to train the networks. For all datasets, we trained the networks using batch size 256 (except for GFDeepCox, IFDeepCox, and DRODeepCox, which have to be trained with full-batch) for 200 epochs. The initial learning rate is set to 0.001, and we decreased it to 0.0001 after the first 100 epochs. Finally, we evaluated the loss function value on the held-out validation set and selected the network parameters that provided the best validation loss.

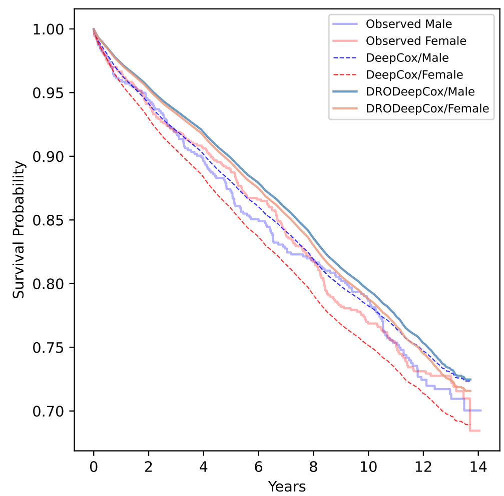
Methods	Hyperparameter	Values
FASTDeepCox FASTDNNSurv FASTDeepHit	γ	[1, 2, 3, 5, 10, 15, 20]
GFDNNSurv IFDNNSurv GFDeepCox IFDeepCox	γ	[0.01, 0.1, 1, 2, 3, 5, 10, 20]
DRODeepCox	η	[1.0, 0.4, 0.3, 0.2, 0.1]

Table A1: The range of the hyperparameters of the fairness-aware models used in the experiment. Note that the grids are *not wide enough* to make the penalty terms dominant because the main purpose of our experiment is to encourage predictive parity as much as possible while not losing the prediction performance.

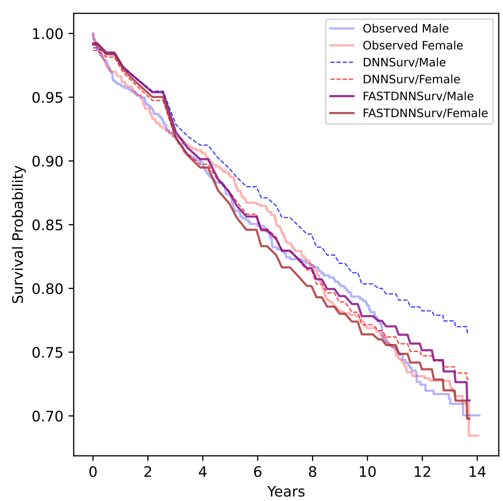
Appendix E. Additional Experimental Results



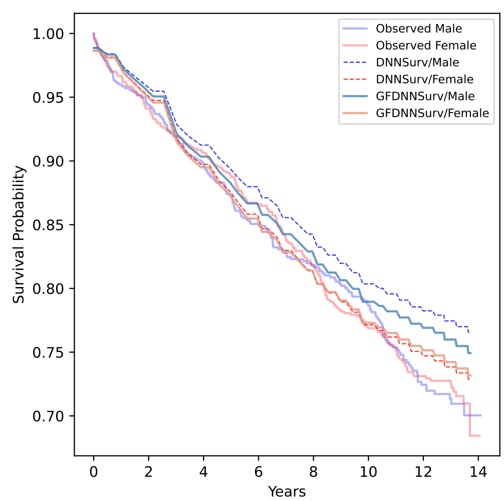
(a) FLChain - GFDeepCox



(b) FLChain - DRODeepCox

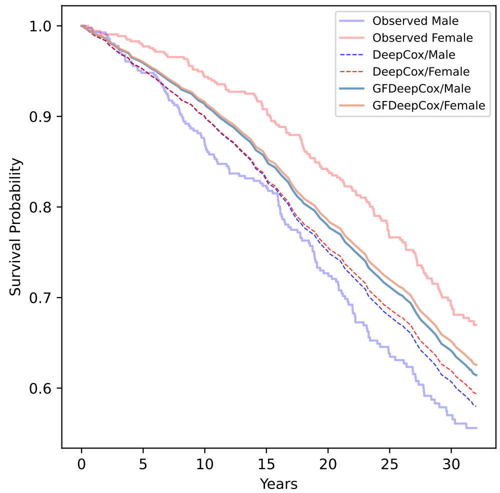


(c) FLChain - FASTDNNSurv

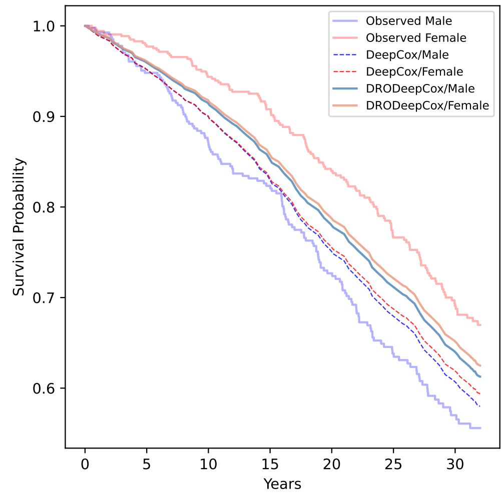


(d) FLChain - GFDNNSurv

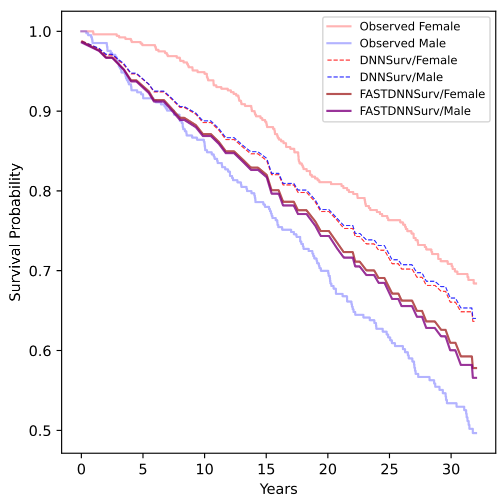
Figure A3: The predicted survival curves generated by DeepCox-based and DNNSurv-based fairness encouraging methods for the FLChain dataset.



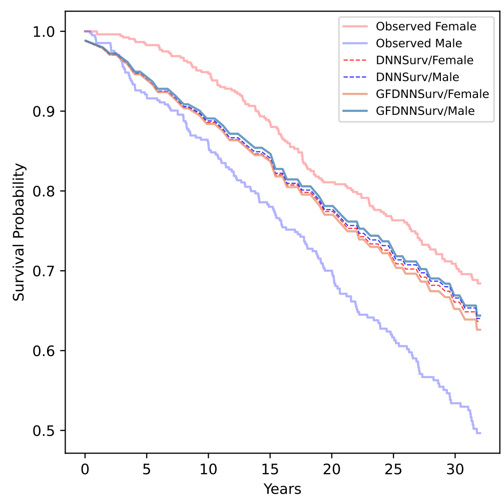
(a) Framingham - GFDeepCox



(b) Framingham - DRODeepCox



(c) Framingham - FASTDNNSurv



(d) Framingham - GFDNNSurv

Figure A4: The predicted survival curves generated by DeepCox-based and DNNSurv-based fairness encouraging methods for the Framingham dataset.

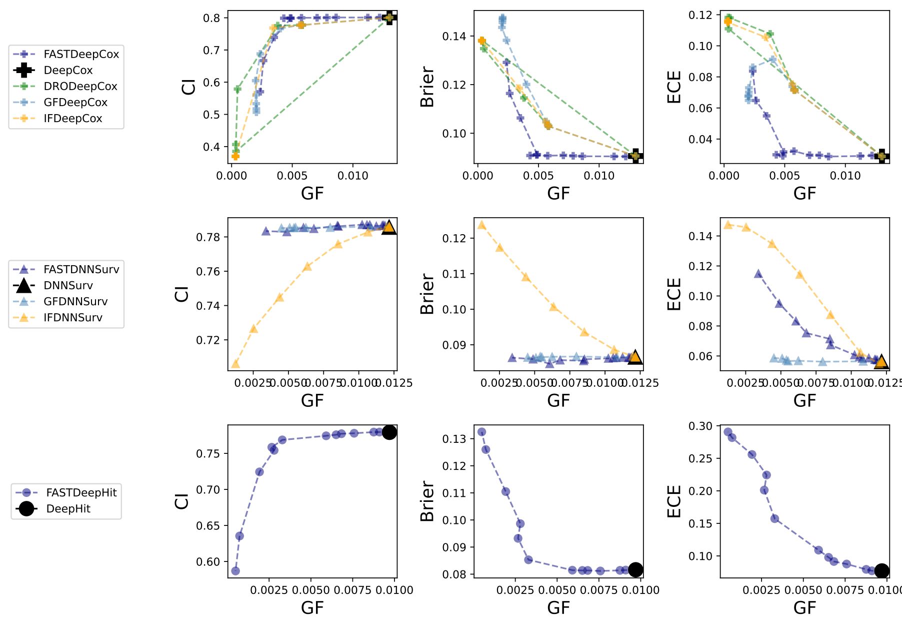


Figure A5: Three performance metrics (Ci, Brier, and Ece) vs. Gf trade-off curves from FLChain dataset.

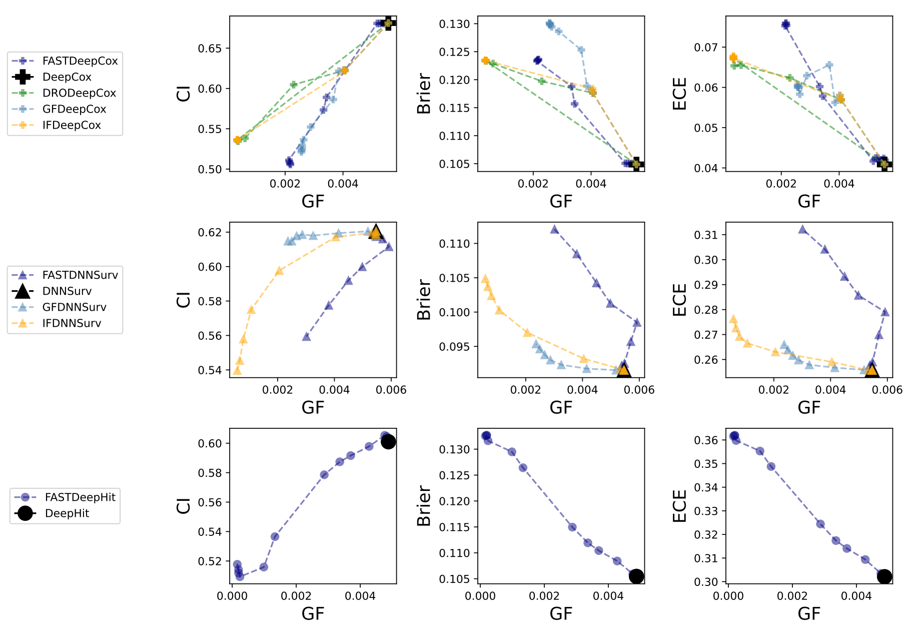


Figure A6: Three performance metrics (Ci, Brier, and Ece) vs. Gf trade-off curves from Framingham dataset.

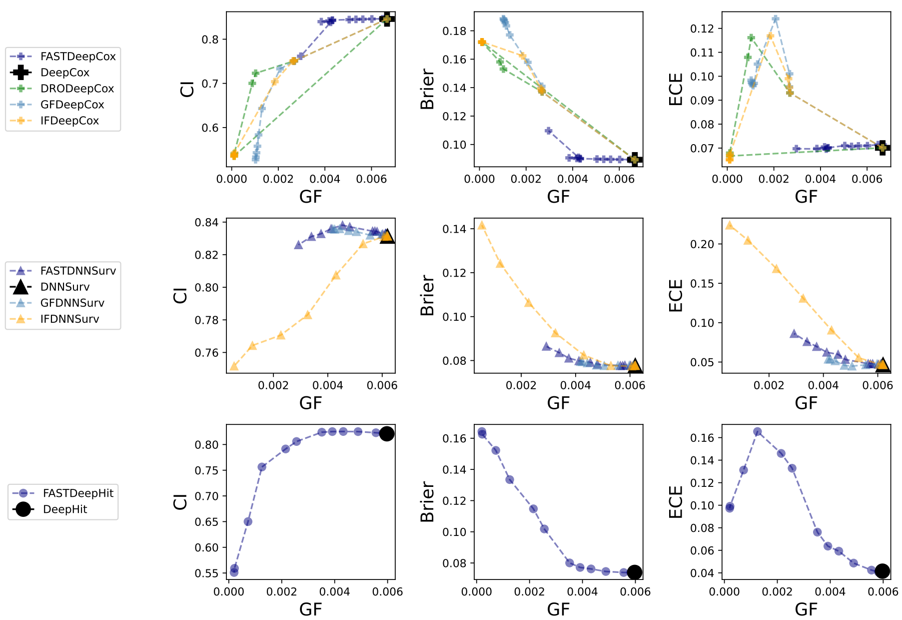


Figure A7: Three performance metrics (Ci, Brier, and Ece) vs. Gf trade-off curves from SUPPORT dataset.

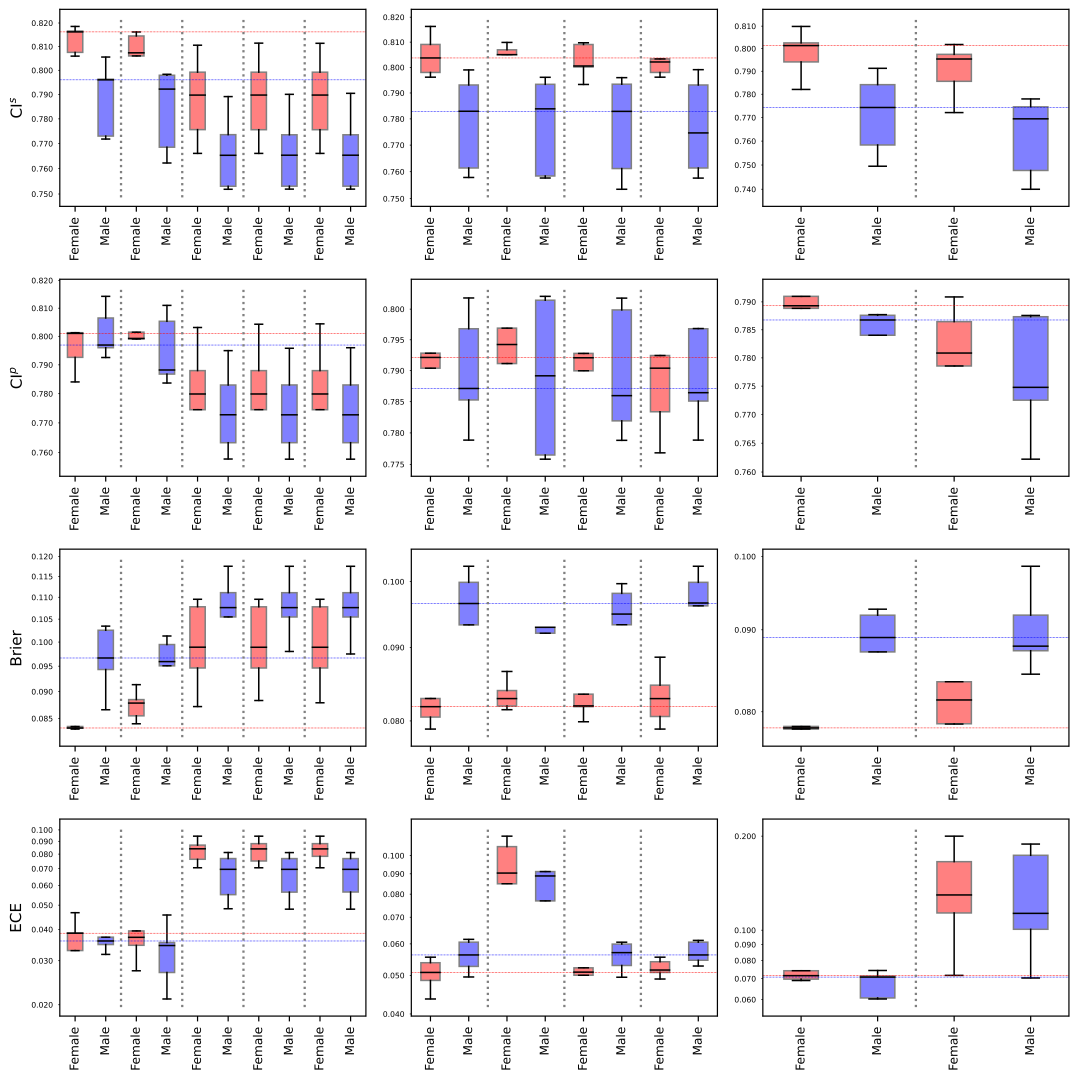


Figure A8: Results of FLChain dataset where the sensitive attribute is gender.

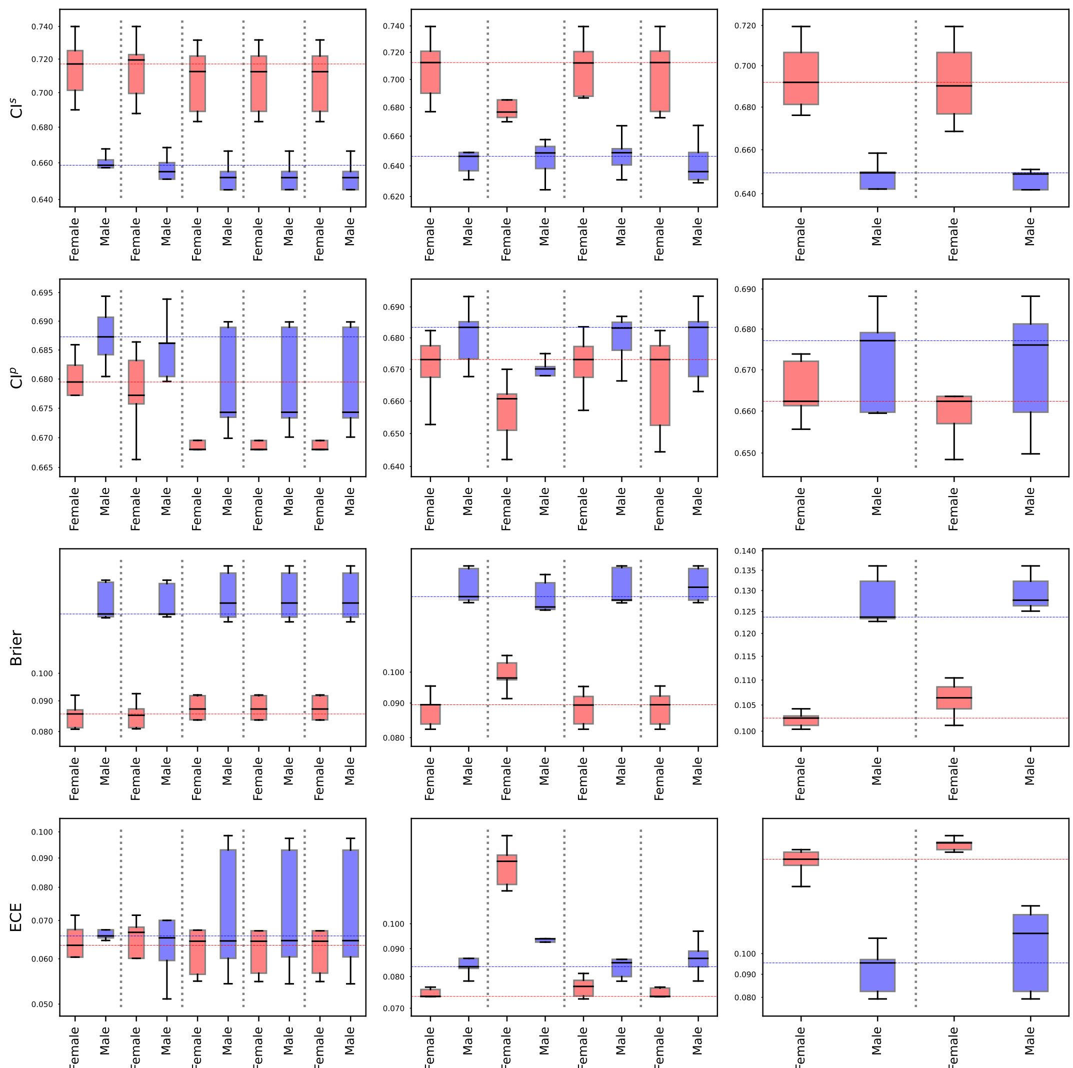


Figure A9: Results of Framingham dataset where the sensitive attribute is gender.

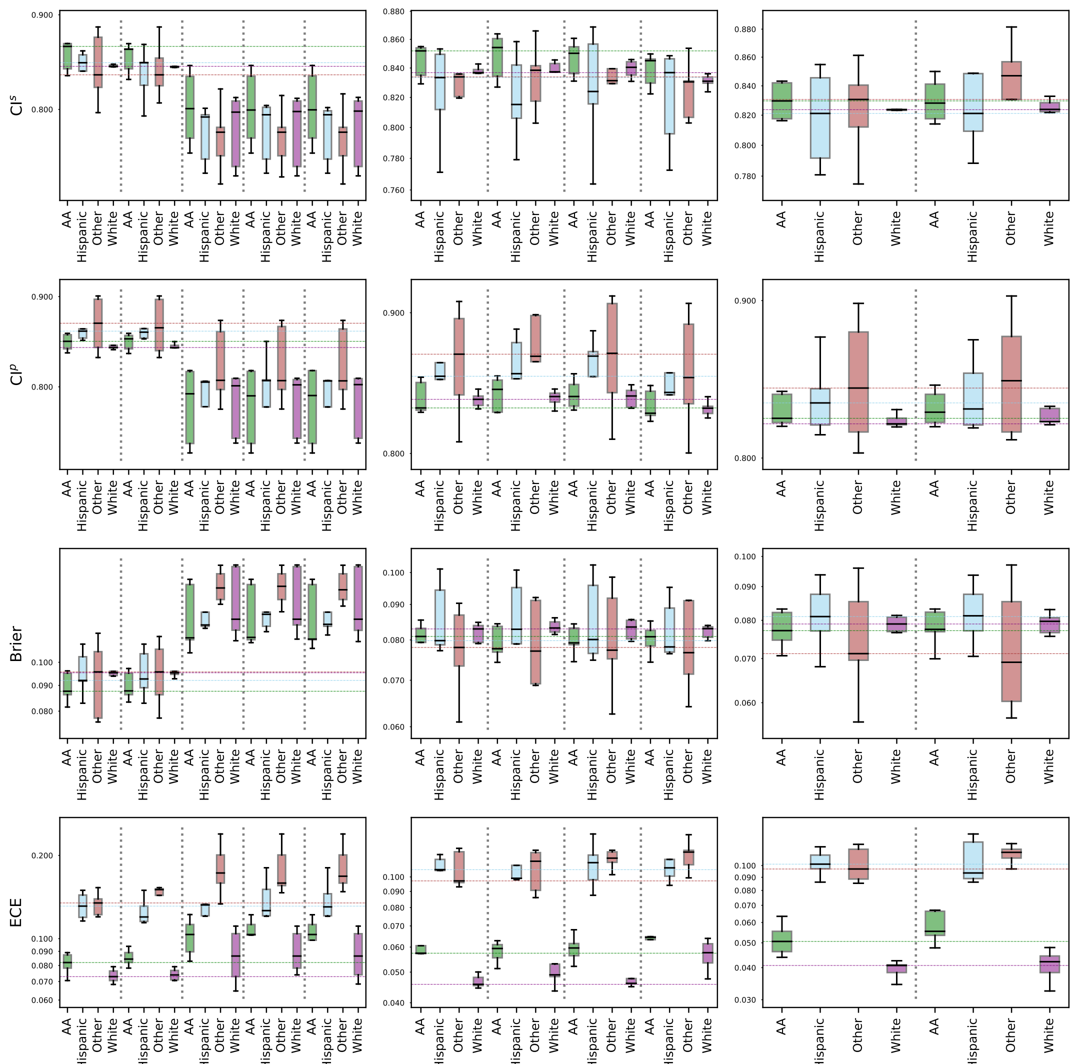
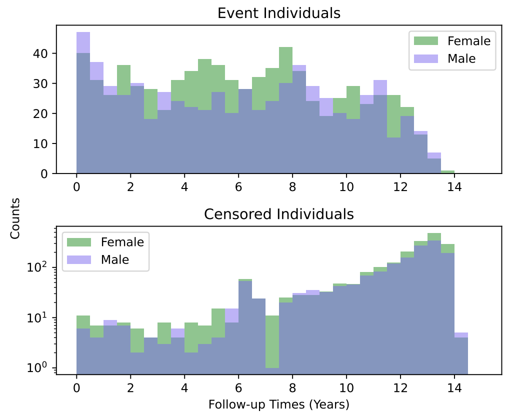
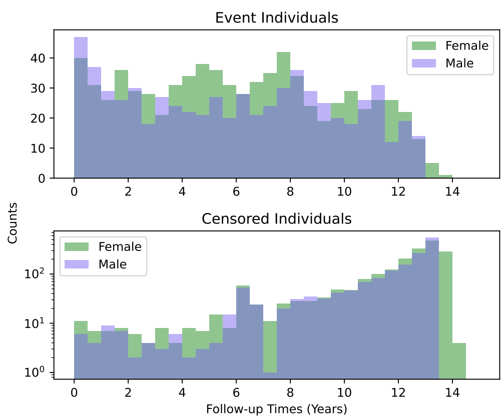


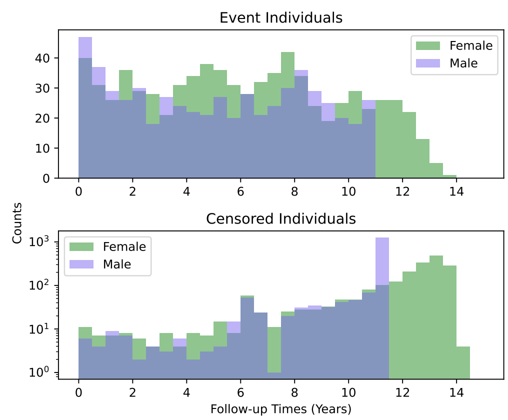
Figure A10: Results of SUPPORT dataset where the sensitive attribute is race/ethnicity.



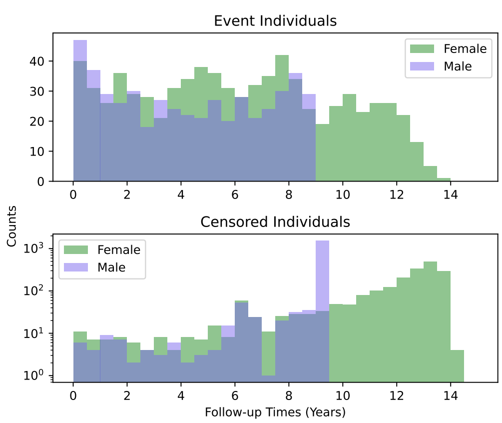
(a) Original Follow-up Times



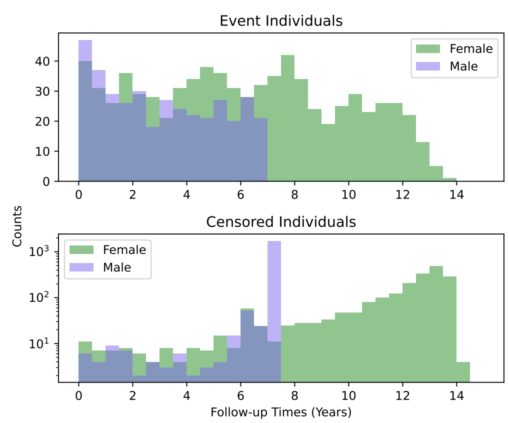
(b) Max Follow-up 13Y



(c) Max Follow-up 11Y

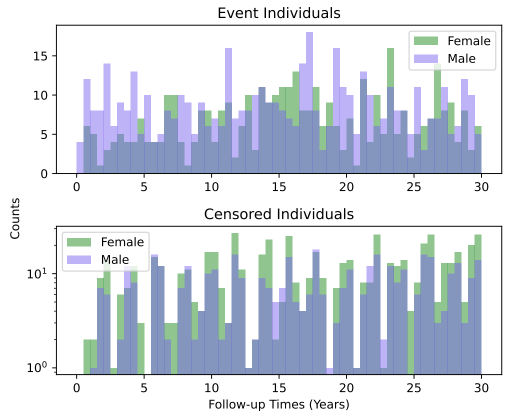


(d) Max Follow-up 9Y

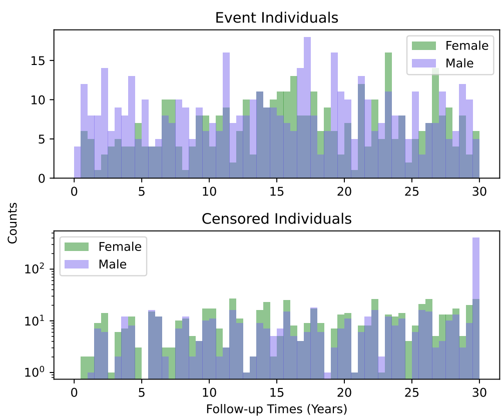


(e) Max Follow-up 7Y

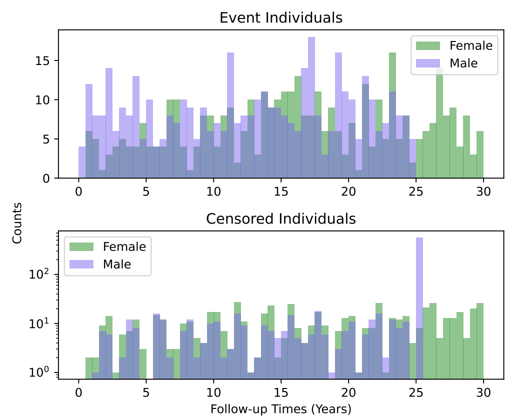
Figure A11: Follow-up time distributions for different max follow-up times for male participants (FLChain).



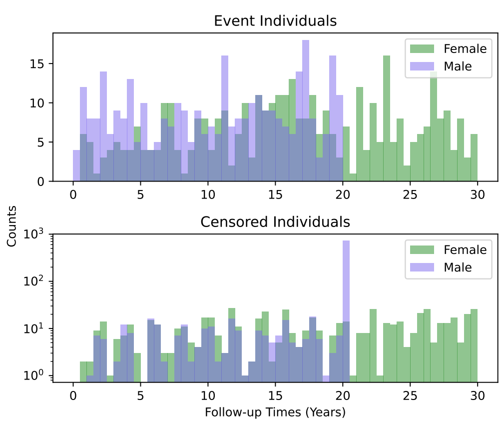
(a) Original Follow-up Times



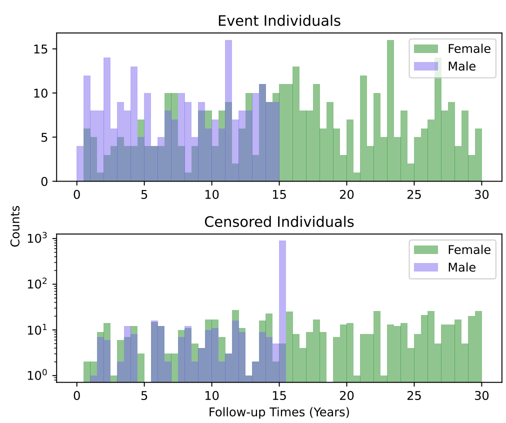
(b) Max Follow-up 30Y



(c) Max Follow-up 25Y



(d) Max Follow-up 20Y



(e) Max Follow-up 15Y

Figure A12: Follow-up time distributions for different max follow-up times for male participants (Framingham).

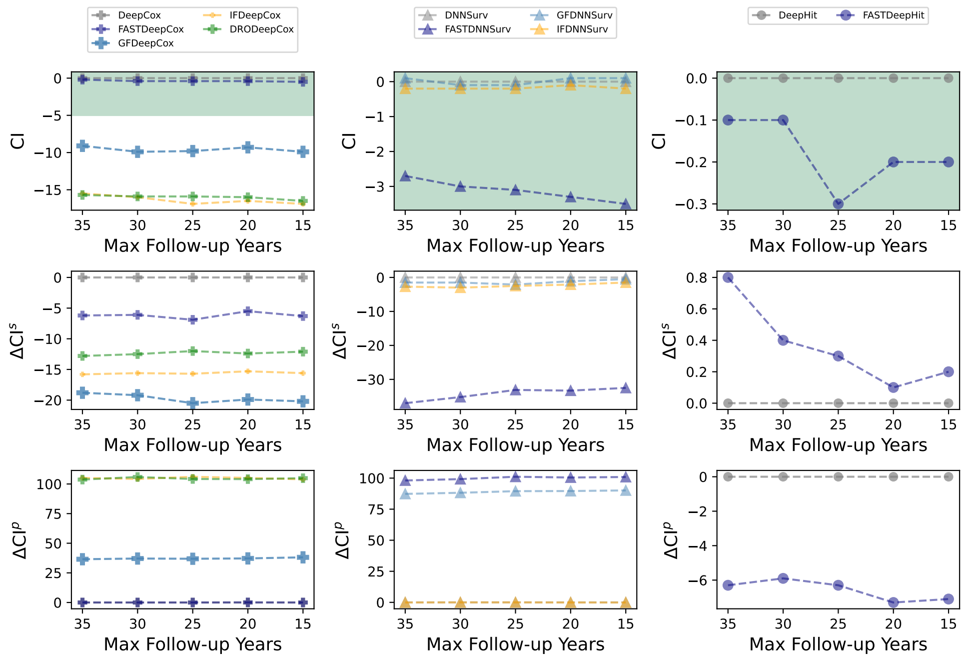


Figure A13: Model performances and disparities versus max follow-up years for males on the Framingham dataset averaged over 5 test sets. The green shades represent the 5% margin of the C-index.

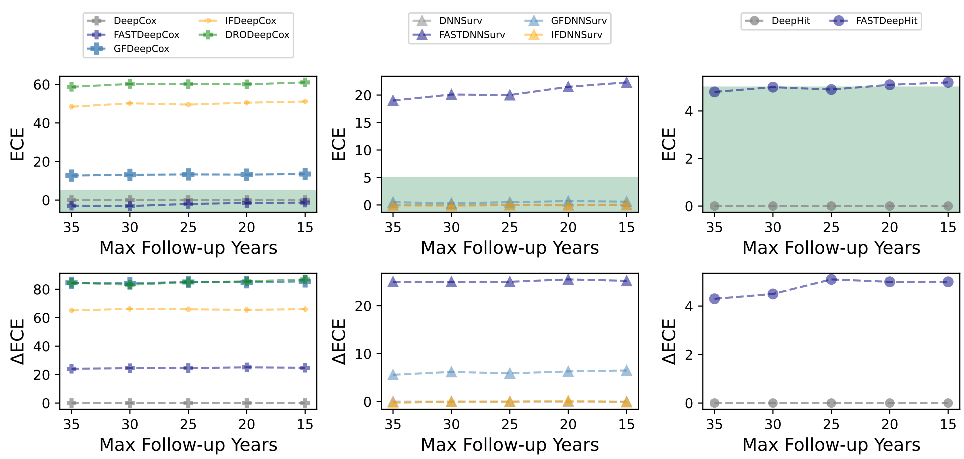


Figure A14: Model performances and disparities versus max follow-up years for males on the Framingham dataset averaged over 5 test sets. The green shades represent the 5% margin of the ECE.

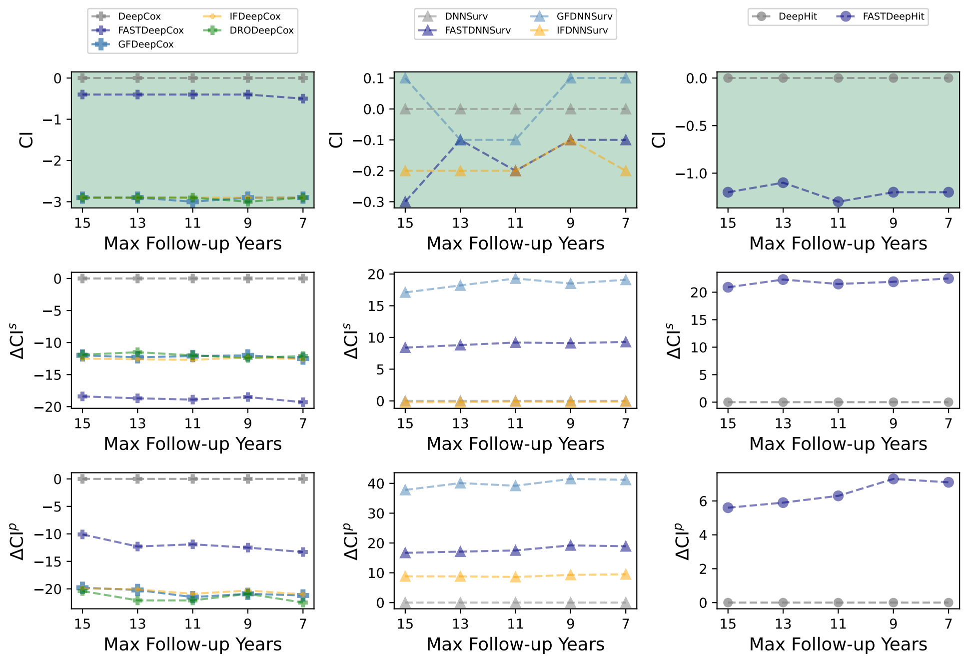


Figure A15: Model performances and disparities versus max follow-up years for males on the FLChain dataset averaged over 5 test sets. The green shades represent the 5% margin of the C-index.

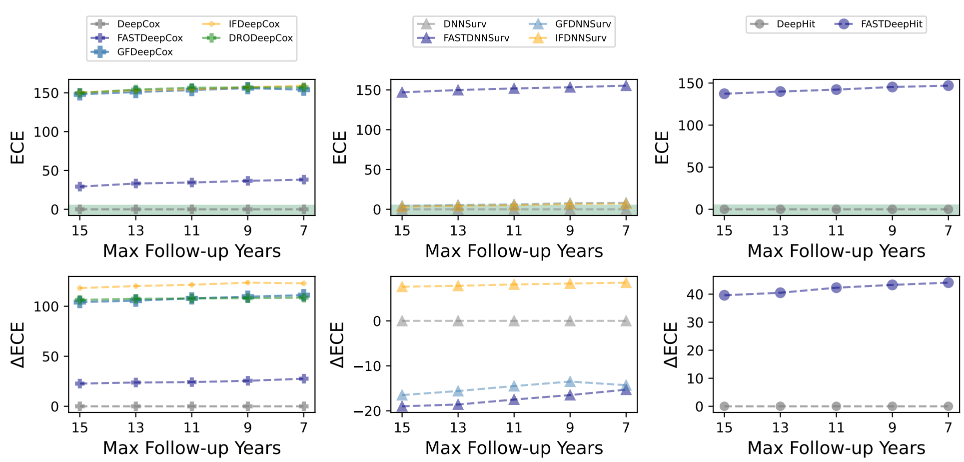


Figure A16: Model performances and disparities versus max follow-up years for males on the FLChain dataset averaged over 5 test sets. The green shades represent the 5% margin of the ECE.

References – Supplementary

Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, 2010.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

Shu Hu and George H. Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Proceedings of the 2nd Machine Learning for Health Symposium*, volume 193, pages 62–87. PMLR, 28 Nov 2022.

Iris AM Huijben, Wouter Kool, Max Benedikt Paulus, and Ruud JG Van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. PMLR, 07–09 Jul 2015.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James Foulds. Equitable allocation of healthcare resources with fair survival models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 190–198. SIAM, 2021.

Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 11074–11101. PMLR, 17–23 Jul 2022.

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

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3384–3393. PMLR, 10–15 Jul 2018.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 807–814. Omnipress, 2010.

Md Mahmudur Rahman and Sanjay Purushotham. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1452–1462, 2022.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2586–2594, 2019.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 20156–20175. PMLR, 17–23 Jul 2022.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *Journal of Machine Learning Research*, 23:1–29, 2022.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 17–19 Jun 2013.

Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3308–3314, 2020.

Qixian Zhong, Jonas W Mueller, and Jane-Ling Wang. Deep extended hazard models for survival analysis. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.