Fair and Interpretable Models for Survival Analysis

Md Mahmudur Rahman
mrahman6@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Sanjay Purushotham
psanjay@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Abstract

Survival analysis aims to predict the risk of an event, such as death due to cancer, in the presence of censoring. Recent research has shown that existing survival techniques are prone to unintentional biases towards protected attributes such as age, race, and/or gender. For example, censoring assumed to be unrelated to the prognosis and covariates (typically violated in real data) often leads to overestimation and biased survival predictions for different protected groups. In order to attenuate harmful bias and ensure fair survival predictions, we introduce fairness definitions based on survival functions and censoring. We propose novel fair and interpretable survival models which use pseudo valued-based objective functions with fairness definitions as constraints for predicting subject-specific survival probabilities. Experiments on three real-world survival datasets demonstrate that our proposed fair survival models show significant improvement over existing survival techniques in terms of accuracy and fairness measures. We show that our proposed models provide fair predictions for protected attributes under different types and amounts of censoring. Furthermore, we study the interplay between interpretability and fairness; and investigate how fairness and censoring impact survival predictions for different protected attributes.

CCS Concepts

• **Mathematics of computing → Survival analysis**; • **Computing methodologies → Neural networks**.

Keywords

Survival analysis; Fairness; Interpretability; Neural networks; Pseudo values; Censoring

ACM Reference Format:

Md Mahmudur Rahman and Sanjay Purushotham. 2022. Fair and Interpretable Models for Survival Analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539259

1 Introduction

Survival analysis, aka time-to-event analysis, is increasingly used for facilitating clinical decision-making [33]. However, recent research works [13, 25, 29] have shown that existing survival analysis techniques are prone to unintentional biases toward protected attributes such as age, race, gender, and/or ethnicity. Moreover, *censoring* (i.e., incomplete information about survival), which leads to an overestimation of survival predictions [5], can more likely occur in a particular demographic group than others [6]. For example, African American women (a minority demographic in the US) experience high exposure to censoring in clinical trials and treatments [30]. In Figure 1, we show the impact of censoring on predicted survival probability curves for different races like White (majority group), Black, Hispanics, and Asians (clustered together as minority group) on the SEER dataset. The left column plot shows that the predicted survival curve (from DNNSurv model [36]) is similar for all demographic groups when the dataset has only uncensored observations. However, when any demographic group (White or non-White) was induced with censored observations - it resulted in significant changes to the survival predictions, especially for survival models without fairness constraints. This example demonstrates the need to handle censoring for protected attributes while building survival models. Thus, in this paper, we investigate the challenges and solutions for achieving fair survival predictions for all the protected groups in the presence of censoring.

Fair survival models were recently proposed by Keya et al. [20], where they used existing fairness definitions [11, 12] as fairness constraints in training Cox-based survival models, CPH [9], and DeepSurv [19]. However, their fair learning algorithms have two main drawbacks: (1) their models use fairness definitions based on hazard functions and thus, are not applicable to the non-hazard-based survival models, and (2) their algorithms do not consider the censoring influence on fairness or survival predictions. To address these issues, in this paper, we propose **pseudo value-based fair deep survival models**, namely **Fair DeepPseudo and Fair PseudoNAM**, which respectively use deep neural networks and neural additive models to provide fair and interpretable survival predictions under various censoring settings. We introduce multiple fairness definitions such as individual fairness, group fairness, and censoring-based fairness definitions, which are defined using survival functions, and thus be used as constraints to make any survival model fair. We model the censored observations using Jackknife-based leave-one out pseudo values [21] and introduce novel pseudo value-based loss functions, which can be designed to handle covariate-dependent censoring. We conducted experiments on three real-world survival datasets to show the fair survival predictions and interpretability obtained by our models in the presence of censoring. Our contributions are:

- We introduce multiple fairness definitions, novel pseudo value-based loss functions, and pseudo valued-based deep survival models with fairness constraints to obtain fair survival predictions for a given protected attribute.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539259

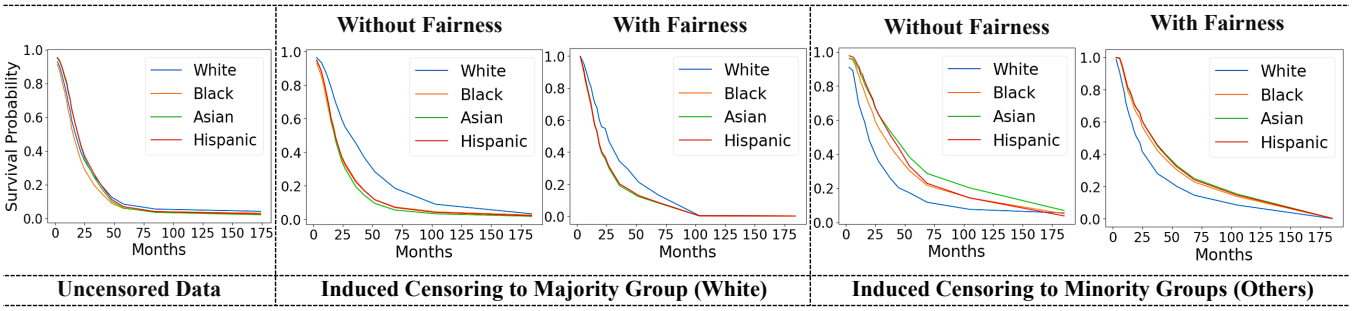

Figure 1: Survival prediction curves for the majority (White) and minority (Black, Asian, Hispanic) groups in the SEER dataset. Survival curves for *without fairness* is obtained from DNNSurv model [36] and survival curves for *with fairness* is obtained from our proposed Fair DeepPseudo (FIDP) model (Section 3.3). In the middle and right column plots, we see that our FIDP model makes fair decisions (i.e., narrows the gap between survival curves of the majority and minority groups) under the induced censoring settings.

- We conduct extensive experiments on three real-world survival datasets to compare and contrast the performance of our proposed models with existing survival models under different amounts and types of censoring and across different protected attributes.
- We investigate the interplay between fairness and interpretability; and also study the impact of fairness and censoring on survival predictions.

2 Related Works

Survival Analysis: We briefly review some of the popular and state-of-the-art survival methods, including Cox Proportional Hazard (CPH) [9], DeepSurv [19], DNNSurv [36], DeepPseudo [27], and PseudoNAM models [28]. Wang et al. [33] provide a detailed survey of statistical and machine learning-based survival models. CPH is a popular semi-parametric survival method that uses hazard functions to estimate the linear effect of covariates on survival risks. DeepSurv is a deep learning-based extension of the CPH model, which learns the nonlinear relationships between the survival risks and covariates. However, both CPH and DeepSurv are limited by proportional hazard assumptions, which may not be satisfied in the real-world data. Moreover, CPH makes a strong linearity assumption. DNNSurv and PseudoNAM are two pseudo value-based deep models for survival analysis. While DNNSurv uses deep neural networks for predicting survival risks, PseudoNAM employs neural additive models for subject-specific interpretable survival predictions. DeepPseudo is a pseudo value-based deep model for competing risk analysis. Although these deep survival models achieve state-of-the-art results, they do not require or guarantee fair survival predictions for protected groups.

Fairness and Bias: Machine learning models are vulnerable to algorithmic bias and prone to making unfair decisions [3, 24]. In order to handle algorithmic bias and circumvent the inequitable and discriminatory predictions from models, researchers have introduced fairness definitions such as individual fairness [11], group fairness [15], intersection fairness [12] and used them as constraints to enforce fairness while training the models [7, 35]. Recent research has shown that bias in the data can adversely impact the analysis and decision making in medical domains [23], which can unfairly affect the patients in minority groups [8]. Thus, there is an urgent need to develop fair learning algorithms for the medical domain.

Fair Survival Analysis: Developing fair survival models is a nascent field, and there are only a few related works [20, 34]. In [20], Keya et al. used existing fairness definitions as regularization terms in the optimization of objective functions of the existing survival models, CPH and DeepSurv. However, their fairness definitions are based on hazard functions and are not applicable for non-hazard survival models. Zheng et al. [34] proposed two censoring-specific fairness notions and a debiasing algorithm based on Random Survival Forests [17] for fair decision making in survival analysis. However, their approach does not address covariate-dependent censoring problems, which may occur in real-world data; and it does not provide interpretable predictions. To overcome the limitations of these works, in this paper, we introduce new fairness definitions based on survival functions and censoring, and propose pseudo value-based fair and interpretable deep survival models.

3 Proposed Fair Survival Models

3.1 Notations

A survival dataset is a collection of time-to-event information about the patients with their corresponding survival status. For a subject i, survival data is a tuple $\{T_i, \delta_i, \mathbf{X_i}\}_{i=1}^{N}$, where, N is total number of subjects, $\mathbf{X_i} = (X_{i1}, ..., X_{ip})$ is a p dimensional vector of observed covariates, T_i is the time until an event or censoring has occurred for subject i, and δ_i is the censoring indicator for subject i. $\delta_i = 1$, if the i^{th} subject experiences the event and $\delta_i = 0$, if the subject is censored. Let $S(t|\mathbf{X})$ represents survival function at time t for the covariate vector $\mathbf{X}$. In fair survival analysis, we are interested in the accurate prediction of $S(t|\mathbf{X})$ for a fairness definition.

3.2 Fairness Definitions

We introduce four different fairness definitions for survival analysis to ensure fair survival predictions for individuals and the protected groups taking into account the censoring.

Individual Fairness: The individual fairness [11] ensures that similar subjects (i.e., subjects with similar characteristics or covariates) have similar outcomes, where similarity is measured using

a distance metric such as a cosine distance or Euclidean distance. Keya et al. [20] introduced individual fairness definitions for hazard-based survival models such as CPH. We generalize their definition to any survival model (hazard or non-hazard based model) and define individual fairness, $F_I(t)$, based on the predicted survival probability ($\hat{S}(t|\mathbf{X})$) at time t, as:

$$F_I(t) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \max(0, |\hat{S}(t|x_i) - \hat{S}(t|x_j)| - \alpha D(x_i, x_j))$$

where $\hat{S}(t|x_k)$ is the predicted survival probability at time t for individual k and $D(x_i, x_j)$ is the distance metric (e.g., cosine distance) between subject i and j's covariates x_i and x_j. α is a scale factor which can be tuned to ensure similar scales for survival probability predictions and cosine distance. Note that our definition measures fairness by computing how much the difference in survival predictions deviates from the cosine distance for similar subjects.

Group Fairness: The group fairness definition [11] warrants that the outcomes across different demographic groups, such as different age groups, gender, or races, are fairly distributed. We define the group fairness, $F_G(t)$, based on the predicted survival probability at time t as:

$$F_G(t) = \max_{a \in A} |E(\hat{S}(t|\mathbf{X}_a)) - E(\hat{S}(t|\mathbf{X}))|$$

where A is the set of values in the protected attribute, and $E(\hat{S}(t|\mathbf{X}_a))$ and $E(\hat{S}(t|\mathbf{X}))$ respectively are the expected predicted survival probabilities for group a and the expected predicted survival probability for the population at time t. This fairness definition measures the maximum deviation of the groups' average survival predictions from the population's average survival predictions.

Censoring-based Individual Fairness: We introduce a censoring based individual fairness definition to capture the potential bias that may arise from censoring due to loss of follow-up or being withdrawn from the study. We define the censoring-based individual fairness, $F_{CI}(t)$ as:

$$F_{CI}(t) = \begin{cases} 0, & \text{if } T_{i,c} > T_{j,uc} \\ \frac{1}{N_c * N_{uc}} \sum\limits_{i \in N_c, j \in N_{uc}} \max(0, |\hat{S}(t|x_{i,c}) - \hat{S}(t|x_{j,uc})| & \\ \qquad -\alpha D(x_{i,c}, x_{j,uc})), & \text{otherwise} \end{cases}$$

where N_c and N_{uc} respectively are the number of censored and uncensored observations, $T_{i,c}$ and $T_{j,uc}$ respectively are the censoring time and survival time of i^{th} and j^{th} subjects, and $\hat{S}(t|x_{i,c})$ and $\hat{S}(t|x_{j,uc})$ are the predicted survival probabilities for i^{th} censored subject and j^{th} uncensored subject, and α is a scale parameter. This fairness definition ensures that a pair of censored and uncensored individuals who have similar covariates have similar survival predictions under the constraint that the censoring time of the censored individual is less than the survival time of the uncensored individual.

Censoring-based Group Fairness: We propose a censoring based group fairness definition to identify the demographic bias which arises from censoring due to loss to follow-up or being withdrawn from the study. Formally, we define the censoring-based group fairness, $F_{CG}(t)$, as:

$$F_{CG}(t) = \begin{cases} 0, & \text{if } T_{i,c} > T_{j,uc} \\ \frac{1}{N_c * N_{uc}} \sum\limits_{g \in G} \sum\limits_{\substack{i \in N_{c,g}, \\ j \in N_{uc,g}}} \max(0, |\hat{S}(t|x^g_{i,c}) - \hat{S}(t|x^g_{j,uc})| & \\ \qquad -\alpha D(x^g_{i,c}, x^g_{j,uc})), & \text{otherwise} \end{cases}$$

where G is the set of groups in a protected attribute, $N_{c,g}$ and $N_{uc,g}$ respectively are the number of censored and uncensored observations in group g (eg. White) of the protected attribute (eg. Race), and $\hat{S}(t|x^g_{i,c})$ and $\hat{S}(t|x^g_{j,uc})$ respectively are the survival probability predictions for i^{th} censored subject and j^{th} uncensored subject from group g. α is a scale parameter. This fairness definition ensures that the censoring based individual fairness holds for each group of the protected attributes.

3.3 Pseudo value-based Fair Survival Models

Pseudo values [2] have been proposed to handle censoring in survival analysis. They can be derived from an asymptotically unbiased estimator such as Kaplan-Meier (KM) estimator [18] for both censored and uncensored observations. Recent research has shown that pseudo value-based deep learning approaches can achieve state-of-the-art results in survival analysis [36] and competing risk analysis [27] without making any underlying assumptions on the stochastic process. While these approaches are useful for obtaining accurate predictions, they do not warrant fair and/or interpretable survival predictions, which are very much needed in the medical domain. Inspired by their success and to address their limitations, we propose two pseudo valued-based fair deep survival models, namely **Fair DeepPseudo** and **Fair PseudoNAM**. Both our proposed models enforce fairness in the learning algorithm by using the fairness definitions (introduced in section 3.2) as a fairness penalty constraint or a regularization term while optimizing a pseudo value-based loss function. Before describing our proposed models, we introduce novel pseudo value-based loss functions and modify them to handle covariate-dependent censoring.

Pseudo value-based Loss Functions: Mean squared error (MSE) loss functions used for training existing pseudo value-based deep survival models such as DNNSurv [36] cannot handle covariate dependent censoring. In heavily censored settings, MSE loss might have convergence issues since the pseudo values and predicted survival probabilities are in different ranges. To address these limitations, we introduce novel pseudo value-based loss functions, which can work for any pseudo-value based survival model. First, we define a pseudo value-based loss function under the covariate independent censoring assumptions and then relax these assumptions to the more general covariate dependent censoring settings.

Let $\hat{S}(t|x_i)$ and $J_i(t)$ be the predicted survival probability and the pseudo values (ground-truth estimated from KM estimator) at time point t for i^{th} individual. Then, we define the pseudo value-based loss function, $L_p(t)$, for training observations N_{train}, at time t as

$$L_p(t) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} [J_i(t)(1 - 2\hat{S}(t|x_i)) + \hat{S}^2(t|x_i)] \qquad (1)$$

under the following assumptions [31]:

Assumption 1: Censoring time $\mathbf{C}$ is independent of the survival time $\mathbf{T}$ and covariates $\mathbf{X}$, i.e. $\mathbf{C} \perp \mathbf{T}, \mathbf{X}$.

Assumption 2: The expectation of the pseudo values for survival probability given the covariates is approximately equal to the actual conditional survival probability given covariates, i.e., $E(J(t)|\mathbf{X}) = S(t|\mathbf{X}) + O_p(1)$.

Pseudo value-based Loss Function with IPCW for Covariate Dependent Censoring: Due to assumption 1, the loss function defined in equation 1 is not suitable for covariate dependent censoring as it leads to biased estimates of the survival probability. In order to get unbiased estimates, we can reweight the pseudo value-based loss function using inverse probability of censoring weight (IPCW) to account for the censored observations given the covariate information. IPCW [31] can be obtained by fitting a consistent model (such as CPH or Random Survival Forests [17]) by assigning censoring time and censoring status as outputs and covariates as inputs. The inverse probability of censoring weights for patient i can be computed as:

$$w_i(t) = 1/\hat{G}(t|x_i)$$

where $\hat{G}(t|X)$ is the estimate of the conditional survival function of the censoring variable given covariates. Thus, for covariate dependent censoring setting, the pseudo value-based loss function with IPCW is defined at time t as:

$$L_{ipcw}(t) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} w_i(t)[J_i(t)(1 - 2\hat{S}(t|x_i)) + \hat{S}^2(t|x_i)] \quad (2)$$

Fair DeepPseudo: Fair DeepPseudo model uses a deep feed forward neural network to directly predict the pseudo values for the survival probability given the covariates as inputs. This model optimizes the pseudo value-based loss function along with a fairness constraint. Thus, the training objective (loss function) is given by:

$$L = L_p + \lambda F_q \quad (3)$$

where L_p is the pseudo value-based loss function and F_q is the fairness constraint obtained from the fairness definitions. $F_q = F_I$ for individual fairness, $F_q = F_G$ for group fairness, $F_q = F_{CI}$ for censoring based individual fairness, and $F_q = F_{CG}$ for censoring based group fairness. The fairness constraint works as regularization term in the objective function. $\lambda > 0$ is the hyperparameter which controls the trade-off between accuracy and fairness.

Fair PseudoNAM: Neural Additive Models (NAM) [1] is a class of neural network models that has the expressivity of deep neural networks and interpretability of generalized additive models. Recently, Rahman et al. [28] proposed adopted NAM models for survival analysis by using pseudo values to handle censoring. Inspired by these works, we propose Fair PseudoNAM, which uses NAM-based model architecture and the pseudo value-based loss function with a fairness constraint to obtain fair and interpretable predictions. Similar to NAM, Fair PseudoNAM consists of a set of separate neural networks for each of the covariates, where each covariate's neural network captures the contribution towards output prediction. Fair PseudoNAM takes the covariates as input and outputs the survival probability at some pre-specified grid of time points. It is interpretable since the non-overlapping neural networks for individual covariates allow the identification of individual covariate's effect on the survival probability prediction. Similar to Fair DeepPseudo, Fair PseudoNAM enforces fairness in the algorithm during training by using the fairness definition as a constraint or a regularization term.

Training Our Proposed Fair Deep Survival Models: Our proposed models are trained by optimizing the pseudo value-based loss function with one of the four fairness definitions as fairness constraints. For individual fairness constraint, the loss function is defined as

$$L = \sum_{t=1}^{T} L_p(t) + \lambda \frac{2 * F_I(t)}{N_{train} * (N_{train} - 1) * T} \quad (4)$$

where $L_p(t)$ is the pseudo value-based loss function at time t and $F_I(t)$ is the individual fairness constraints at time t. λ is the tradeoff parameter between accruacy and fairness, T is the number of pre-specified prediction time points, N_{train} is number of observations in the training data. For censoring based fairness penalty, the loss function is defined as

$$L = \sum_{t=1}^{T} L_p(t) + \lambda \frac{F_C(t)}{T} \quad (5)$$

$F_C(t)$ can be $F_{CI}(t)$ or $F_{CG}(t)$.

Fair DeepPseudo and Fair PseudoNAM models with pseudo-value loss function L_p and individual fairness constraints are respectively denoted as **FIDP** and **FIPNAM**. For covariate dependent censoring setting, $L_p(t)$ is replaced with $L_{ipcw}(t)$; and the corresponding Fair DeepPseudo and Fair PseudoNAM models are denoted as **FIDPipcw** and **FIPNAMipcw** respectively.

4 Experiments

Empirically, we answer these research questions: (a) how do our fair survival models compare to the existing survival approaches? (b) how do our fair survival models perform under different censoring settings? (c) what is the effect of fairness and censoring on survival predictions? (d) how does fairness impact interpretability?

4.1 Datasets

We conducted experiments on the following three real-world survival datasets under varying censoring settings. Additional details about these datasets are provided in the appendix A.1.

FLChain: The FLChain dataset [10] is collected from a study of the relationship between serum free light chain (FLC) and mortality of the Olmsted County residents aged 50 years or more. The preprocessed dataset contains 6521 individuals and 8 covariates and is highly censored (70% of the observations are censored). The median survival time for the uncensored patients is 2084 days, and the median censoring time for the censored patients is 4621 days. Two of the covariates are protected attributes - *age* and *gender*.

SUPPORT: The SUPPORT dataset [22] is obtained from a Vanderbilt University study, which was conducted to understand patient survival for seriously ill hospitalized patients. It has 8950 patients including 2863(32%) censored patients and 32 covariates. The median survival time for the uncensored patients is 58 days and median censoring time for the censored patients is 916 days. Three of its covariates - *age*, *gender* and *race* are protected attributes.

SEER: The Surveillance, Epidemiology, and End Results (SEER)[1] Program of National Cancer Institute provides information on the survival attributes of oncology patients in the United States. The dataset contains 28366 breast cancer patients, out of which 75% are

[1] https://seer.cancer.gov/

Table 1: Model performance comparisons in terms of accuracy measures and individual fairness measures. Higher (↑) Cindex and AUC; and lower (↓) values of Brier score, individual fairness F_I, censoring based individual fairness F_{CI} are better.

Model	FLChain					SUPPORT					SEER				
	Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓	Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓	Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓
CPH [9]	**0.785**	0.149	0.811	0.257	0.030	0.742	0.193	0.812	0.249	0.026	0.806	0.153	0.843	0.271	0.213
DCPH [19]	0.784	0.149	0.811	0.259	0.030	0.742	0.193	0.811	0.254	0.025	0.807	0.152	0.844	0.260	0.219
DNNSurv [36]	0.782	0.154	0.807	0.275	0.029	0.761	**0.182**	**0.842**	0.287	0.017	**0.818**	**0.145**	**0.858**	0.255	0.084
PseudoNAM [28]	0.764	0.160	0.787	0.237	0.025	0.704	0.191	0.822	0.268	0.018	0.799	0.153	0.836	0.240	0.081
FCPH [20]	0.784	0.149	0.811	0.257	0.030	0.741	0.192	0.810	0.251	0.026	0.808	0.151	0.846	0.272	0.212
FDCPH [20]	0.784	0.149	0.810	0.259	0.030	0.740	0.194	0.809	0.251	0.026	0.807	0.152	0.846	0.255	0.231
FIDP	0.773	**0.137**	0.799	0.164	0.017	0.765	0.190	0.837	0.205	**0.009**	0.805	0.165	0.847	**0.129**	0.044
FIPNAM	0.768	0.153	0.798	0.189	0.022	0.715	0.188	0.831	0.208	0.012	0.797	0.155	0.835	0.208	0.069
FIDPipcw	0.765	0.462	0.808	**0.124**	**0.012**	**0.766**	0.217	0.839	0.206	**0.009**	0.779	0.474	0.817	**0.129**	**0.039**
FIPNAMipcw	0.759	0.485	0.787	0.129	0.013	0.690	0.218	0.813	**0.179**	0.010	0.748	0.427	0.828	0.138	0.043

Table 2: Model comparisons based on group fairness measures: censoring based group fairness F_{CG} and group fairness on protected attributes age $F(G_{age})$, gender $F(G_{gender})$ and race $F(G_{race})$. Lower fairness value indicates better performance.

Model	FLChain			SUPPORT			SEER		
	F_{CG} ↓	$F(G_{age})$ ↓	$F(G_{gender})$ ↓	F_{CG} ↓	$F(G_{age})$ ↓	$F(G_{race})$ ↓	F_{CG} ↓	$F(G_{age})$ ↓	$F(G_{race})$ ↓
CPH [9]	0.016	0.209	0.003	0.016	0.041	0.015	0.072	0.195	0.112
DCPH [19]	0.016	0.210	0.003	0.016	0.044	0.013	0.074	0.184	0.097
DNNSurv [36]	0.015	0.220	**0.002**	0.010	0.045	0.014	0.029	0.175	0.086
PseudoNAM [28]	0.013	0.158	0.015	0.011	0.058	0.015	0.027	0.185	0.095
FCPH [20]	0.016	0.208	**0.002**	0.016	0.044	0.012	0.072	0.198	0.102
FDCPH [20]	0.016	0.210	0.006	0.016	0.043	0.016	0.078	0.168	0.105
FIDP	0.009	0.122	**0.002**	**0.005**	0.028	0.014	**0.014**	0.121	0.065
FIPNAM	0.011	0.139	0.022	0.007	0.031	0.011	0.023	0.173	0.107
FIDPipcw	**0.006**	**0.090**	0.005	0.006	**0.021**	0.009	**0.013**	**0.113**	**0.033**
FIPNAMipcw	**0.007**	**0.089**	0.010	0.007	0.024	**0.004**	0.015	0.139	0.048

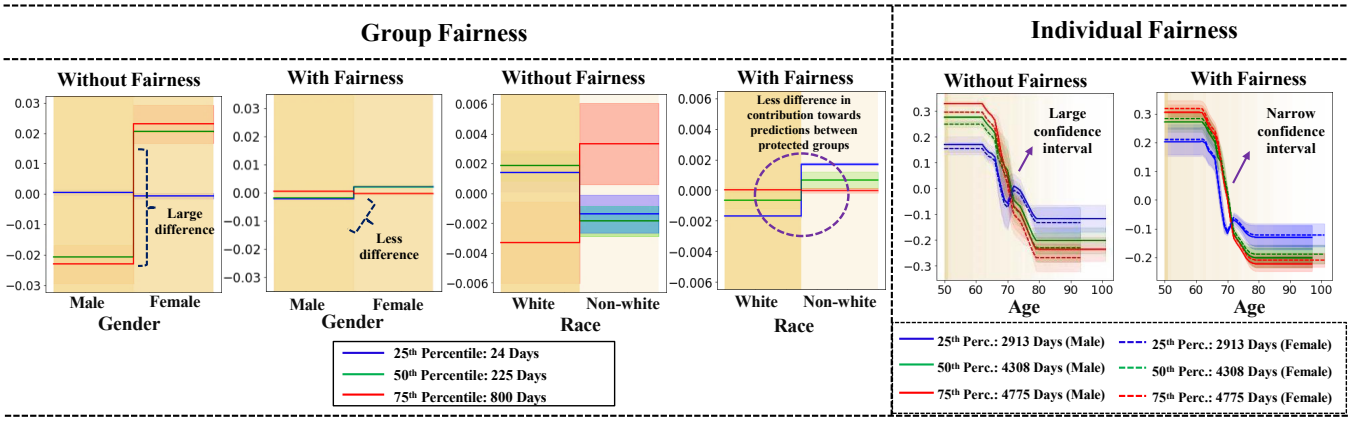

Figure 2: Interpretability of the protected attributes' contribution to the survival probability predictions before and after applying fairness constraints on SUPPORT data (Group fairness) and FLChain data (Individual fairness) using *Fair PseudoNAM* model. y-axis: the protected attributes' contribution to the survival probability predictions. x-axis: protected attribute values.

censored, and it includes 13 covariates. The median survival time of the patients who experienced cancer due to breast cancer is 18 months, and the median censoring time is 80 months. It contains two protected attributes - *age* and *race*.

4.2 Model Comparisons

We compare the following survival models: (1) Cox Proportional Hazard Model (**CPH**) [9], (2) DeepSurv (**DCPH**) [19], (3) **DNNSurv** [36], (4) **PseudoNAM** [28], (5) Fair CPH (**FCPH**) [20], (6) Fair Deep

Table 3: Comparison of the models on different censoring settings with respect to group fairness measures. Lower values indicate better model performance. Our proposed models are: FIDP, FIPNAM, FIDPipcw, and FIPNAMipcw.

Censoring Setting	Model	FLChain			SUPPORT			SEER		
		$F_{CG}\downarrow$	$F(G_{age})\downarrow$	$F(G_{gender})\downarrow$	$F_{CG}\downarrow$	$F(G_{age})\downarrow$	$F(G_{race})\downarrow$	$F_{CG}\downarrow$	$F(G_{age})\downarrow$	$F(G_{race})\downarrow$
Increment All	**FCPH [20]**	0.014	0.230	0.021	0.015	0.059	0.010	0.036	0.134	0.017
	FDCPH [20]	0.015	0.218	0.015	0.014	0.059	0.014	0.037	0.125	0.021
	FIDP	0.007	0.145	0.018	0.007	0.045	0.010	0.013	0.105	**0.011**
	FIPNAM	0.009	0.182	0.008	0.007	0.047	0.007	0.015	0.140	0.023
	FIDPipcw	0.006	0.124	0.008	**0.003**	0.050	0.007	**0.010**	0.094	0.013
	FIPNAMipcw	**0.005**	**0.102**	**0.005**	0.004	**0.017**	**0.004**	0.011	**0.083**	0.020
Increment Majority	**FCPH [20]**	0.015	0.178	0.198	0.008	0.008	0.141	0.014	0.099	0.188
	FDCPH [20]	0.015	0.172	0.198	0.009	0.013	0.129	0.013	0.075	0.191
	FIDP	0.008	0.112	0.139	0.005	0.019	0.110	0.009	0.064	0.120
	FIPNAM	0.009	0.103	0.176	0.004	0.005	0.107	0.008	0.071	0.116
	FIDPipcw	**0.007**	**0.091**	**0.123**	**0.003**	**0.002**	**0.093**	**0.006**	**0.070**	**0.096**
	FIPNAMipcw	0.008	0.111	0.137	0.004	0.006	0.096	0.007	0.093	0.112
Increment Minority	**FCPH [20]**	0.012	0.185	0.189	0.015	0.038	0.117	0.030	0.074	0.259
	FDCPH [20]	0.012	0.175	0.191	0.016	0.042	0.118	0.028	0.072	0.273
	FIDP	0.008	0.155	0.172	0.010	0.013	**0.071**	0.013	0.056	0.208
	FIPNAM	0.007	0.155	0.168	0.012	0.035	0.090	0.011	**0.052**	0.126
	FIDPipcw	0.007	0.131	**0.120**	**0.004**	**0.010**	0.076	**0.008**	0.053	**0.101**
	FIPNAMipcw	**0.003**	**0.066**	0.130	0.004	0.014	0.088	0.008	0.063	0.144
Induced Majority	**FCPH [20]**	0.052	0.049	0.088	0.126	0.032	0.083	0.025	0.021	0.149
	FDCPH [20]	0.049	0.054	0.096	0.122	0.032	0.092	0.026	0.009	0.144
	FIDP	**0.004**	0.023	**0.046**	0.072	0.025	0.058	0.008	0.055	0.115
	FIPNAM	0.009	0.018	0.063	0.049	0.022	0.059	0.011	0.020	0.055
	FIDPipcw	0.007	0.026	0.046	**0.031**	**0.017**	**0.027**	0.009	0.034	0.020
	FIPNAMipcw	0.006	**0.007**	0.048	0.039	0.020	0.048	**0.009**	**0.018**	**0.015**
Induced Minority	**FCPH [20]**	0.055	0.131	0.086	0.053	0.008	0.063	0.038	0.068	0.153
	FDCPH [20]	0.052	0.142	0.096	0.053	0.010	0.066	0.037	0.043	0.157
	FIDP	0.018	0.085	0.050	0.038	0.027	0.074	**0.009**	**0.028**	0.086
	FIPNAM	0.014	0.063	0.067	0.033	0.008	0.066	0.012	0.038	0.069
	FIDPipcw	0.016	0.084	0.031	**0.029**	0.020	0.040	0.012	0.067	0.065
	FIPNAMipcw	**0.012**	**0.051**	**0.028**	0.030	**0.006**	**0.026**	0.016	0.052	**0.055**
Uncensored	**FCPH [20]**	0.000	0.051	0.013	0.000	**0.003**	**0.007**	0.000	0.070	0.056
	FDCPH [20]	0.000	0.050	0.018	0.000	0.004	0.007	0.000	0.067	0.048
	FIDP	0.000	0.040	0.005	0.000	0.028	0.019	0.000	**0.054**	**0.043**
	FIPNAM	0.000	**0.006**	**0.003**	0.000	0.040	0.013	0.000	0.067	0.054

Cox Proportional Hazards (**FDCPH**) [20] with our proposed models: **FIDP**, **FIPNAM**, **FIDPipcw**, and **FIPNAMipcw**.

4.3 Implementation and Evaluation Metrics

We construct the training (64%), validation (16%), and test (20%) data by stratifying the data based on protected attributes and censoring. For model performance comparisons, we use **accuracy measures** such as (a) time-dependent concordance index (**Cindex**) [4], (b) integrated IPCW Brier Score (**Brier**) [14] and (c) mean cumulative AUC (**AUC**) [32], and **fairness measures** such as (a) individual fairness (F_I), (b) censoring based individual fairness (F_{CI}), (c) group fairness (F_G) and (d) censoring based group fairness (F_{CG}) as the evaluation metrics. Hyperparameter tuning was done on the validation data. More details about the implementation and pseudo code for our fair algorithms are provided in the Appendix A.4. The source code is available at https://github.com/umbc-sanjaylab/FISA_KDD22.

4.4 Censoring Settings

We conduct a detailed investigation on how the model's performance and fairness are impacted by varying types and amounts of censoring on a protected attribute. We consider six censoring settings based on the **incremental censoring** and **induced censoring** mechanisms [27] and on the protected attributes *Race* and *Gender*. Incremental censoring enables us to study the impact of censoring on uncensored observations in an increasing sized dataset, while induced censoring helps us to study the effect of increasing censoring ratio in a fixed-sized dataset. We studied the following censoring settings: (1) Uncensored: Contains an equal number of uncensored subjects from each protected group (SEER: 500 each from White, Black, Asian and Hispanic; SUPPORT: 1000 each from White and Non-white; FLChain: 500 each from Male and Female); (2) Increment All: Add s censored observations ($s = 500$) from the censored cohort of the original dataset to the *uncensored setting*. This setting helps to study the effect of equally adding censoring to the groups (e.g., White, non-White) of the protected attribute (e.g., Race); (3) Increment Majority: Add s censored observations ($s = 500$) only to the majority group (e.g., White) from the censored cohort of the original dataset; (4) Increment Minority: Add s censored observations ($s = 500$) to all the minority groups only from the censored cohort of the original dataset; (5) Induced Majority: Induce r censored observations ($r = 250$ for SEER and FLChain, 500 for SUPPORT) from the uncensored observations for majority group only. Thus, minority groups have the same number of uncensored observations as in the *uncensored setting*; (6) Induced Minority: Induce r censored observations ($r = 250$ for SEER and FLChain, 500

Table 4: Comparison of models on different censoring settings in terms of accuracy measures and individual fairness measures. Higher (↑) Cindex and AUC, and lower (↓) values of Brier score, (F_I) and (F_{CI}) are better.

Censoring Setting	Model	FLChain					SUPPORT					SEER				
		Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓	Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓	Cindex ↑	Brier ↓	AUC ↑	F_I ↓	F_{CI} ↓
Increment All	FCPH [20]	0.785	0.223	0.853	0.299	0.028	0.756	0.193	0.839	0.289	0.017	0.752	0.201	0.800	0.277	0.145
	FDCPH [20]	0.787	0.224	0.855	0.283	0.029	0.755	0.196	0.838	0.301	0.017	0.752	0.198	0.802	0.268	0.149
	FIDP	0.782	0.250	0.855	0.204	0.014	**0.780**	**0.181**	**0.863**	0.265	0.008	**0.762**	0.197	**0.814**	0.205	0.053
	FIPNAM	0.760	**0.212**	0.810	0.208	0.018	0.689	0.212	0.804	**0.199**	**0.007**	0.757	**0.192**	0.806	0.227	0.060
	FIDPipcw	**0.785**	0.367	**0.855**	0.170	0.012	0.775	0.206	0.855	0.246	0.007	0.745	0.233	0.801	**0.160**	**0.041**
	FIPNAMipcw	0.758	0.373	0.833	**0.121**	**0.010**	0.714	0.220	0.851	0.217	0.008	0.715	0.226	0.788	0.169	0.044
Increment Majority	FCPH [20]	0.715	0.259	0.783	0.257	0.044	0.770	0.148	0.855	0.283	0.017	0.732	0.141	0.792	0.279	0.091
	FDCPH [20]	0.716	0.260	0.784	0.251	0.044	0.767	0.149	0.851	0.266	0.018	0.730	0.142	0.791	0.282	0.090
	FIDP	0.725	**0.253**	0.724	0.191	0.023	**0.775**	0.152	0.866	0.297	0.009	**0.736**	**0.119**	0.799	0.190	0.047
	FIPNAM	**0.726**	0.263	0.786	0.211	0.026	0.731	**0.143**	**0.869**	0.280	0.009	0.725	0.124	0.799	0.178	0.037
	FIDPipcw	0.721	0.343	**0.792**	**0.161**	**0.020**	0.774	0.155	0.855	0.278	**0.005**	0.706	0.139	**0.805**	**0.177**	**0.033**
	FIPNAMipcw	0.715	0.372	0.788	0.170	0.022	0.717	0.146	0.859	**0.239**	0.008	0.731	0.121	0.792	0.188	0.039
Increment Minority	FCPH [20]	**0.741**	0.261	**0.800**	0.255	0.033	0.770	0.150	0.861	0.280	0.016	0.731	0.180	0.788	0.257	0.126
	FDCPH [20]	0.740	0.261	0.799	0.243	0.034	0.768	0.149	0.858	0.272	0.017	0.733	0.184	0.789	0.275	0.120
	FIDP	0.729	**0.256**	0.788	0.217	0.024	0.786	0.147	0.870	0.285	0.010	**0.743**	**0.175**	**0.801**	0.241	0.054
	FIPNAM	0.712	0.300	0.787	0.203	0.022	0.733	**0.146**	0.870	0.267	0.011	0.710	0.188	0.775	0.196	0.044
	FIDPipcw	0.733	0.346	0.793	**0.171**	0.019	**0.786**	0.151	**0.880**	0.244	**0.008**	0.736	0.205	0.795	**0.148**	**0.031**
	FIPNAMipcw	0.695	0.332	0.774	0.140	**0.012**	0.716	0.153	0.859	**0.230**	0.009	0.714	0.219	0.784	0.161	0.035
Induced Majority	FCPH [20]	0.585	0.277	0.572	0.141	0.238	**0.790**	0.176	0.876	0.280	0.198	**0.660**	0.146	0.714	0.193	0.267
	FDCPH [20]	0.581	0.277	0.567	0.159	0.232	0.790	0.185	0.876	0.293	0.192	0.656	0.144	0.710	0.182	0.269
	FIDP	0.563	0.277	0.572	0.078	**0.034**	0.767	0.161	0.873	0.280	0.109	0.656	0.139	**0.715**	0.180	0.089
	FIPNAM	0.588	**0.272**	0.580	0.092	0.050	0.740	0.196	0.844	**0.184**	**0.076**	0.626	0.146	0.701	0.120	0.067
	FIDPipcw	0.564	0.311	0.568	0.074	0.038	0.788	**0.133**	**0.883**	0.210	0.087	0.642	**0.134**	0.712	**0.074**	0.061
	FIPNAMipcw	**0.589**	0.322	**0.598**	**0.065**	0.039	0.743	0.145	0.849	0.220	0.102	0.536	0.145	0.620	0.086	**0.060**
Induced Minority	FCPH [20]	0.625	0.254	0.656	0.187	0.206	0.730	0.198	0.811	0.247	0.212	**0.681**	0.230	**0.715**	0.191	0.230
	FDCPH [20]	0.625	0.255	0.655	0.200	0.204	0.723	0.196	0.801	0.238	0.217	0.677	0.236	0.712	0.195	0.225
	FIDP	0.642	0.259	**0.686**	0.149	0.063	**0.737**	0.198	**0.819**	0.264	0.111	0.671	0.214	0.706	**0.116**	**0.055**
	FIPNAM	**0.643**	**0.247**	0.670	0.128	0.051	0.710	**0.164**	0.802	0.209	0.092	0.638	0.200	0.677	0.122	0.063
	FIDPipcw	0.630	0.310	0.662	0.128	0.053	0.730	0.153	0.807	0.196	0.087	0.646	0.170	0.685	0.145	0.070
	FIPNAMipcw	0.630	0.301	0.643	**0.090**	**0.037**	0.699	0.140	0.792	**0.163**	**0.082**	0.615	**0.164**	0.660	0.155	0.081
Uncensored	FCPH [20]	0.583	**0.272**	0.612	0.090	0.000	0.755	0.129	0.855	0.265	0.000	0.641	0.148	0.705	0.185	0.000
	FDCPH [20]	0.576	0.273	0.605	0.101	0.000	0.755	**0.125**	0.855	**0.250**	0.000	0.641	0.151	0.704	0.193	0.000
	FIDP	**0.598**	0.277	**0.634**	0.105	0.000	**0.768**	0.133	0.863	0.309	0.000	**0.645**	**0.125**	**0.705**	**0.140**	0.000
	FIPNAM	0.533	0.282	0.498	**0.046**	0.000	0.715	0.143	**0.876**	0.293	0.000	0.625	0.130	0.684	0.160	0.000

for SUPPORT) from the uncensored observations for all the minority groups (e.g., non-White). Thus, the majority group (e.g., White) has the same number of uncensored observations as in the *uncensored setting*. Note: censoring is induced by flipping the event status of the uncensored subjects. Settings (3) and (4) help us to understand the censoring effect on the accuracy and fairness measures of majority & minority groups, respectively, due to incremental censoring; while settings (5) and (6) help us to understand the effect of increasing censoring ratio on the accuracy and fairness measures of majority & minority groups for a fixed-sized dataset.

5 Results and Discussion

Model Performance Comparisons: Table 1 shows the model performance comparison results in terms of accuracy measures (Cindex, Brier, AUC) and individual fairness measures (F_I, F_{CI}), and Table 2 shows all the models' results in terms of group fairness measures ($F_{CG}, F(G_{age}), F(G_{gender}), F(G_{race})$) for the three datasets. From these tables, we have the following observations: (1) our proposed models (FIDP, FIPNAM, FIDPipcw and FIPNAMipcw) perform significantly better than all other survival models with respect to all fairness measures, (2) Our proposed models obtained similar or sometimes slightly better performance in terms of Cindex and AUC accuracy measures and slightly worse performance in terms of Brier score, (3) We note that state-of-the-art fair survival models [20] surprisingly perform similar to Cox-based survival models (CPH, DCPH); (4) all the pseudo value-based survival models (DNNSurv, PseudoNAM, and our proposed models) generally outperform other survival models (which don't use pseudo values) in terms of accuracy and fairness measures.

Performance under Different Censoring Settings: Tables 3 and 4 show the performance of fair survival models under various censoring settings (as described in section 4.4) in terms of accuracy and fairness measures. From these tables, we see that our proposed models ((FIDP, FIPNAM, FIDPipcw and FIPNAMipcw) significantly outperform the other fair models (FCPH, FDCPH) in almost every case under different censoring settings. This indicates that our proposed models, which use pseudo values and pseudo valued-based loss functions, are less sensitive to censoring and provide fair and

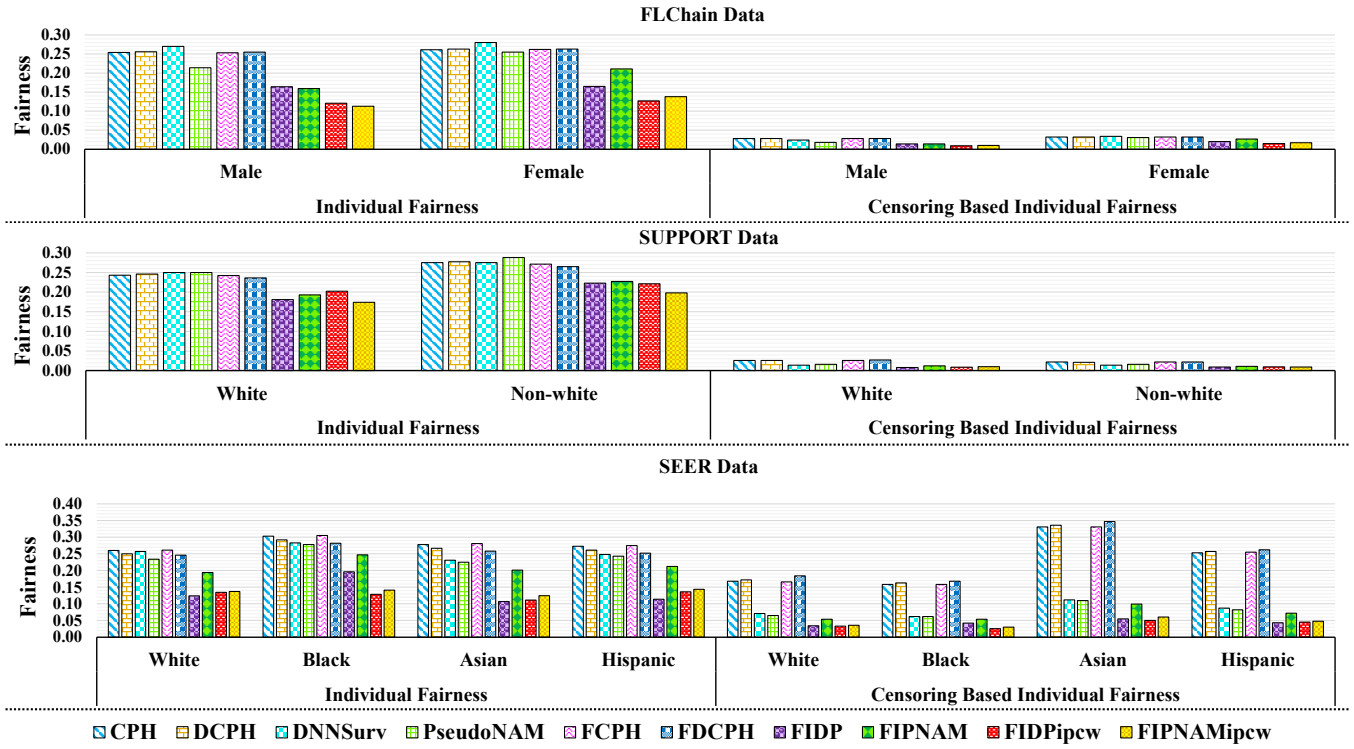

Figure 3: Comparison of the performance and fairness of the models across the protected groups. The lower value of fairness measure indicates better performance. The last four bars (purple, green, red and yellow colors) are our proposed models.

accurate predictions in the presence of different amounts and types of censoring.

Model Comparisons across the Protected Attributes: Figure 3 shows the fairness measures (F_I, F_{CI}) for all the survival models across protected attributes for the three datasets. Results are for *gender* protected attribute for the FLChain data, *race* protected attribute for the SEER and SUPPORT datasets. From this figure, we have two main takeaways: (1) our proposed models (shown as purple, green, red, and yellow colors bars) show better fairness scores (lower fairness value is better) compared to all other survival models across the protected groups; (2) Our proposed models achieve similar fairness values across protected groups which imply that minority groups (e.g., non-White) in the protected attributes (e.g., Race) are equally benefited as the majority groups (e.g., White). For example, in the SEER dataset, the bar heights for censoring-based individual fairness (F_{CI}) are similar for White, Black, Asian, and Hispanic groups.

Impact of Fairness and Censoring on the Survival Predictions: We study the impact of fairness and censoring on survival predictions by (1) varying censoring to the protected groups (Majority and Minority groups) and (2) obtaining the average survival prediction curves with a fair survival model (with fairness) and other regular survival models (without fairness) on the SEER dataset. In particular, we induce censored observations to the majority (White) and minority (Black, Asian, Hispanic) groups respectively from the uncensored data by flipping the event status (from 1 to 0) and reducing the survival time by a random amount to reproduce the loss to follow-up censoring. Figure 1 shows the average survival prediction curves for uncensored (left column plot), induced censoring to the majority group (middle column plot), and induced censoring to the minority group (right column plot) using DNNSurv model (without fairness) and our FIDP model (with fairness). In this figure, we notice that the average survival prediction curves are similar for all the protected groups. However, when censored observations are induced to a particular group, the DNNSurv model (without fairness) results in an elevated survival prediction curve for the censored group (for example: in the middle column plot, the White group has induced censoring, and the survival curve (blue line) is higher and further away from the survival curves of the non-White groups). However, our FIDP model (with fairness) forces the survival prediction curves of censored and uncensored groups to be closer to each other due to the fairness constraint. Thus, this plot clearly shows that our proposed model reduces bias (which could be introduced due to censoring) and makes the model fair towards the protected groups. Thus, our proposed models achieve good survival predictions while ensuring fair predictions for all the protected groups.

Fairness and Interpretability: In Figure 2, we demonstrate the interplay between fairness and interpretability by plotting and visualizing the covariate contribution toward survival probability predictions. In particular, this figure shows the interpretability of

the protected attributes' contribution toward the survival probability predictions. The plots in the left column show the NAM shape functions learned on the SUPPORT data by PseudoNAM (without fairness) and our Fair PseudoNAM (FIPNAM) model (with fairness) for *gender* and *Race* protected attribute under group fairness constraint. NAM shape functions shown here were learned by an ensemble of 20 NAMs. The plots in the right column show the NAM shape functions learned on the FLChain data by PseudoNAM (without fairness) and Fair PseudoNAM model (with fairness) for *Age* and *Race* protected attributes under individual fairness constraints. In both these plots, we show contributions for three different prediction time points - 25^{th}, 50^{th}, 75^{th} percentile of the survival times in the training data. The left plot clearly shows that in our model FIPNAM, the covariate contribution of the protected group is almost similar, i.e., the model is fair (in terms of group fairness) w.r.t the survival probability prediction. In other words, both male and female groups have the same impact on the survival prediction. While PseudoNAM model (without fairness) does not impose group fairness, and thus the two groups (male and female) contribute different amounts towards survival probability prediction. We observe similar visualization insights in the left column plot for *Race* protected attribute, where our FIPNAM model, which enforces fairness, has similar covariate contributions towards model survival predictions at all the prediction time points.

In the right column plot of figure 2, we see that the survival probability tends to decrease with the increase in age for both males and females. Survival probability at all three time points sharply drops after 65 years and keeps decreasing up to 80 years. After 80 years, the survival probability becomes low and stable. At an earlier prediction time (25^{th}: 2913 days), survival probability slightly increased after 70 years and continued up to 75 years, and then again decreased. From this right plot, we can conclude that in our model FIPNAM (with fairness), the contribution of age and gender towards the survival probability predictions slightly changes, and the confidence intervals in the survival probability prediction period are narrower than for the PseudoNAM model (without fairness). This indicates that FIPNAM (with fairness) more confidently captures the change in the covariates' contribution to the survival probability than PseudoNAM (without fairness) models.

6 Conclusions

In this paper, we introduce multiple fairness definitions for survival analysis in the presence of censoring, suitable for any survival model. We utilized these fairness definitions to develop fair survival learning algorithms by using them as constraints while optimizing a novel pseudo value-based loss function. We empirically showed that our proposed fair survival models are less sensitive to different amounts and types of censoring. We graphically demonstrated that our models improve the fairness and interpretability of the survival predictions across different protected groups. In our future work, we will investigate the appropriateness and advantages of developing fair learning algorithms with censoring-based fairness constraints.

Acknowledgement

This work is supported by grant IIS–1948399 from the US National Science Foundation (NSF).

References

[1] Rishabh Agarwal et al. 2020. Neural additive models: Interpretable machine learning with neural nets. *arXiv:2004.13912* (2020).
[2] Per Kragh Andersen et al. 2010. Pseudo-observations in survival analysis. *Statistical methods in medical research* (2010).
[3] Julia Angwin et al. 2019. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. (2019).
[4] Laura Antolini et al. 2005. A time-dependent discrimination index for survival data. *Statistics in medicine* (2005).
[5] Enrique Barrajon et al. 2020. Effect of right censoring bias on survival analysis. *arXiv preprint arXiv:2012.08649* (2020).
[6] J Martin Bland et al. 2004. The logrank test. *BMJ* (2004).
[7] Tolga Bolukbasi et al. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS* (2016).
[8] Irene Y Chen et al. 2019. Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics* (2019).
[9] David R Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* (1972).
[10] Angela Dispenzieri et al. 2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*. Elsevier.
[11] Cynthia Dwork et al. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*.
[12] James R Foulds et al. 2020. An intersectional definition of fairness. In *ICDE*.
[13] Milena A Gianfrancesco et al. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* (2018).
[14] Erika Graf et al. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* (1999).
[15] Moritz Hardt et al. 2016. Equality of opportunity in supervised learning. *NeurIPS* (2016).
[16] Frank E Harrell et al. 1982. Evaluating the yield of medical tests. *JAMA* (1982).
[17] Hemant Ishwaran et al. 2008. Random survival forests. *The annals of applied statistics* (2008).
[18] Edward L Kaplan et al. 1958. Nonparametric estimation from incomplete observations. *JASA* (1958).
[19] Jared L Katzman et al. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology* (2018).
[20] Kamrun Naher Keya et al. 2021. Equitable Allocation of Healthcare Resources with Fair Survival Models. In *SDM*. SIAM.
[21] John P Klein et al. 2008. SAS and R functions to compute pseudo-values for censored data regression. *Computer methods and programs in biomedicine* (2008).
[22] William A Knaus et al. 1995. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine* (1995).
[23] Arjun K Manrai et al. 2016. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine* (2016).
[24] Ninareh Mehrabi et al. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* (2021).
[25] Vishwali Mhasawade et al. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence* (2021).
[26] Chirag Nagpal et al. 2021. Deep Cox mixtures for survival regression. *arXiv preprint arXiv:2101.06536* (2021).
[27] Md Mahmudur Rahman et al. 2021. DeepPseudo: Pseudo Value Based Deep Learning Models for Competing Risk Analysis. In *AAAI*.
[28] Md Mahmudur Rahman et al. 2021. PseudoNAM: A Pseudo Value Based Interpretable Neural Additive Model for Survival Analysis. *AAAI FSS* (2021).
[29] Alvin Rajkomar et al. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* (2018).
[30] Kevin A Schulman et al. 1999. The effect of race and sex on physicians' recommendations for cardiac catheterization. *NEJM* (1999).
[31] Cristian Spitoni et al. 2018. Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal* (2018).
[32] Hajime Uno et al. 2007. Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* (2007).
[33] Ping Wang et al. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)* (2019).
[34] Wenbin Zhang et al. 2021. Fair Decision-making Under Uncertainty. In *ICDM*.
[35] Jieyu Zhao et al. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP* (2017).
[36] Lili Zhao et al. 2020. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics* (2020).

A Appendix

A.1 Dataset details

Table 5 shows the descriptive statistics of three real-world datasets - SEER, SUPPORT, FLCHAIN datasets. Table 6 shows the summary statistics for SEER and SUPPORT datasets based on protected attribute *race*.

A.2 Implementation Details

We first divide the datasets into censored and uncensored groups, and then divide each censored and uncensored group into subgroups of the protected attributes, such as `race`. We hold out 20% as the test set from each subgroup and choose the rest 80% as the training set. Furthermore, we set aside 20% from the training set as a validation set. Then we combine the train, validation, and test sets from each subgroup to get the final train, validation, and test sets. This ensures the equal distribution of observations in each subgroup and an equal censoring ratio. We obtain the pseudo values (ground-truth) for survival probability using the 'jackknife' function of R package 'prodlim' for M evaluation time points (separately for training and validation sets). We optimize the models using Adam optimizer with early stopping criteria based on validation loss. For the models without fairness, we run the models for 5000 epochs with patience 50 with an early stopping criteria, and for the models with fairness, we set 100 epochs with patience 10 with an early stopping criteria. The Fair DeepPseudo model consists of 5 hidden layers with a dropout of 0.4 following each layer. We use `SELU` activation in the hidden layers and the `sigmoid` activation function in the output layer to obtain the survival probability. In the Fair PseudoNAM model, each neural network corresponding to the covariates consists of 3 hidden layers with a number of units [128, 64, 32]. We use `ReLU` activation function in the hidden layer of each covariate's neural network and `sigmoid` activation function in the output layer to get the survival probability. We set the batch size of 128 and the learning rate of 0.01 for both of our fair models. For performance metrics, we used (a) time-dependent concordance index [4], (b) Integrated IPCW Brier Score [14] and (c) Mean cumulative AUC [32]. For the fairness measures, we consider (a) individual fairness, (b) censoring-based individual fairness, (c) group fairness, and (d) censoring-based group fairness measure. We ran experiments on a 128GB RAM Intel Xeon dual 10-core processor with 3 GPUs.

A.3 Tuning Fairness-Accuracy Trade-off Parameter λ

The performance of the models can be impacted by the fairness parameter λ. λ should be selected in a way that achieves the best fairness with a minimum accuracy loss. In our experiments, we trained the models by varying the values of λ for different scale parameters: 0.1, 0.01, and 0.001. We selected the parameter λ for which we achieved overall better performance with respect to C-index, Brier Score, and AUC score while achieving fair results on the validation data. In this paper, we set the trade-off parameter $\lambda = 0.1$ and scale parameter $\alpha = 0.01$ for all the models after hyperparameter tuning on the validation data.

A.4 Pseudo code for our proposed fair models

Algorithm 1 provides the pseudo-code to help implement our fair learning algorithms. The source code is publicly available at https://github.com/umbc-sanjaylab/FISA_KDD22.

Algorithm 1: Fair DeepPseudo, Fair PseudoNAM

Input : Covariates $\mathbf{X}$, N_{train} subjects, Pre-specified time points $t = 1, ..T$, Fairness constraints F_q.

Output : Predicted survival probabilities $\hat{\mathbf{S}}(\mathbf{t}|\mathbf{X})$, and model weights $\mathbf{\Phi}$.

for $i \leftarrow 1$ **to** N_{train} **do**
  Pseudo values $J_i \leftarrow$ KM estimator using Jackknife leave-one-out approach.
end for
if *Covariate independent censoring* **then**
  Compute
  $L_p(t) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} [J_i(t)(1 - 2\hat{S}(t|x_i)) + \hat{S}^2(t|x_i)]$
else
  Compute (covariate dependent censoring): $L_p(t) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} w_i * [J_i(t)(1 - 2\hat{S}(t|x_i)) + \hat{S}^2(t|x_i)]$
end if
Optimize $L = L_p(t) + \lambda F_q$, where $F_q = F_I$ or F_{CI} or F_{CG} or F_G to learn model weights $\mathbf{\Phi}$.
return $\hat{\mathbf{S}}(\mathbf{t}|\mathbf{X})$ *and* $\mathbf{\Phi}$

A.5 Evaluation Metrics

We evaluate the models with respect to time-dependent concordance index (C-index) [4], integrated Brier Score (Brier score) [14] and mean cumulative or dynamic area under ROC Curve (AUC) score [32].

Time-dependent Concordance Index: Concordance index (C-index) is rank order statistics that measures the ratio of concordant pairs to comparable pairs. Comparable pair refers to two instances where both of them are uncensored or the observed event time of the uncensored instance is smaller than the censoring time of the censored instance. A pair is said to be a concordant pair if the instance whose survival time is less has less survival probability than the other instance. We use time-dependent concordance index (C^{td}) [4], which compares the relative risks of all pairs in the test set at fixed evaluation time horizons to measure the ranking ability of the model.

$$C^{td} = P\{\hat{P}(T > t|X_i) < \hat{P}(T > t|X_j)|\delta_i = 1, T_i < T_j, T_i \leq t\}$$

C^{td} captures possible changes in risk over time and relax the constant proportional hazards made by C-index [16].

Brier Score: The Brier score is a proper scoring rule for both discriminative performance and calibration of a model's estimates. The Brier score computes the Mean Squared Error around the binary outcomes from survival data at a fixed time t.

$$BS(t) = E[(\mathbb{I}(T > t) - \hat{P}(T > t|X))^2]$$

Graf et al. [14] adjusted the Brier score for censoring using inverse probability of censoring weighting (IPCW) as:

$$BS_{ipcw}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\hat{P}(T > t|X_i)^2 \mathbb{I}\{T \leq t, \delta_i = 1\}}{\hat{G}_i(T_i)} + \frac{(1 - \hat{P}(T > t|X_i))^2 \mathbb{I}\{T > t\}}{\hat{G}_i(t)} \right]$$

Here, $\hat{G}(.)$ is the Kaplan Meier estimate of the Censoring Distribution. BS_{ipcw} is an unbiased estimate of the Brier Score when

Table 5: Descriptive statistics of three real-world survival datasets used in our experiments

Dataset	N	No. of Features or Covariates	Censoring(%)	Event Quantiles			Censoring Quantiles		
				t=25th	t=50th	t=75	t=25th	t=50th	t=75
FLChain	6521	8	70%	907.5	2084.0	3245.0	4122	4621	4855
SUPPORT	8950	32	32%	14	58	253	622	916	1533.5
SEER	25319	13	75%	10	18	35	29	80	145

Table 6: Summary statistics of the datasets based on protected attribute *Race*

Dataset	Race	N (%)	Censoring (%)	Event Quantiles			Censoring Quantiles		
				t=25th	t=50th	t=75	t=25th	t=50th	t=75
SUPPORT	White	7190 (80%)	31%	14	60	259	622	917	1547
	Black	1391 (16%)	35%	13	49	226	608	902.5	1511.5
	Asian	79 (1%)	27%	7.25	41.5	157.25	727	864	1599
	Hispanic	290 (3%)	40%	18	59	229	649	958	1430
SEER	White	13807 (55%)	75%	10	19	37	33	86	150
	Black	2889 (11%)	65%	9	16	29	29	81	146
	Asian	6207 (25%)	79%	11	19	34	22	67	131
	Hispanic	2416 (10%)	76%	11	19	35	29	81	152

the censoring times and survival times are independent. We use **integrated Brier score**, which can computed by extending the BS from single duration t to an interval as: $IBS = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} BS(s)ds$.

Area under ROC Curve (AUC): The area under the ROC curve for survival analysis is computed by treating the survival analysis problem as binary classification at different quantiles of event times [26]. The cumulative or dynamic AUC quantifies the discriminative ability of a model in distinguishing subjects who experience an event by a given time ($t_i \leq t$) from subjects who experience an event after this time ($t_i > t$). The cumulative or dynamic AUC is adjusted by inverse probability of censoring weights (IPCW) as proposed in [32]. We use the mean cumulative or dynamic AUC that can be computed by integrating the cumulative or dynamic AUC over the time range (τ_1, τ_2).