Censored Fairness through Awareness

Wenbin Zhang[1], **Tina Hernandez-Boussard**[2] and **Jeremy Weiss**[3]

[1]Michigan Technological University, Houghton, MI 49931
[2]Stanford University, Stanford, CA 94305
[3]National Institutes of Health, Bethesda, MD 20892
wenbinzh@mtu.edu, boussard@stanford.edu, jeremy.weiss@nih.gov

Abstract

There has been increasing concern within the machine learning community and beyond that Artificial Intelligence (AI) faces a bias and discrimination crisis which needs AI fairness with urgency. As many have begun to work on this problem, most existing work depends on the availability of class label for the given fairness definition and algorithm which may not align with real-world usage. In this work, we study an AI fairness problem that stems from the gap between the design of a "fair" model in the lab and its deployment in the real-world. Specifically, we consider defining and mitigating individual unfairness amidst censorship, where the availability of class label is not always guaranteed due to censorship, which is broadly applicable in a diversity of real-world socially sensitive applications. We show that our method is able to quantify and mitigate individual unfairness in the presence of censorship across three benchmark tasks, which provides the first known results on individual fairness guarantee in analysis of censored data.

Introduction

AI-based decision-making systems, when implemented in real-life scenarios, have been shown to exhibit bias and discrimination against marginalized groups or populations. This is evidenced by instances in various fields, such as criminal justice (Chouldechova 2017), healthcare (Chen et al. 2020), predictive policing (Chang 2021), and employment (Miller 2015). As a result, there is a growing body of research on quantifying and guaranteeing fairness for machine learning (Beutel et al. 2019; Meyer 2018; Skirpan and Gorelick 2017). The vast majority of them address the problem by taking the statistical group fairness approach that first identifies a small collection of high-level groups defined by the *sensitive attribute*, such as gender or race, then ensures similar outcome statistics of the predictor (*e.g.*, the prediction accuracy and true positive rate), across these groups, with the aim of preventing practices that one socially salient group is collectively allocated a more favorable outcome (*e.g.*, which patients need extra medical care and the targeted customers to receive promotional deals) compared to another (Mehrabi et al. 2021; Zhang et al. 2021; Saxena, Zhang, and Shahabi 2023a,b). In addition, another common theme amongst all these prior works is the assumption of fairness as a supervised learning problem– where the class label either actual or predicted is given as a precondition for fairness definitions as well as debiasing algorithms depending on these notions to enforce fairness (Žliobaitė 2017; Zhang and Ntoutsi 2019; Quy et al. 2022).

This setting, however, is unrealistic to many domains leaving users with real-world socially sensitive problems without tooling to mitigate discrimination and prejudice concerns– echoing existing critiques that current fairness-aware methods do not meet the real-world fair AI use cases (Hoffmann 2019; Selbst et al. 2019). In this work, we study such an AI fairness problem that originates from the gap between the design of a "fair" model in the lab and their real-world deployment. Specifically, we consider the ubiquitous *censorship* phenomenon in real-world data analysis in which the assumption of class label guarantee does not hold, but still requires that similar individuals are treated similarly (Dwork et al. 2012). Below exemplifies such a real-life AI fairness problem that necessitates censorship management,

Example 1. A hospital plans to create precise AI algorithm to help streamline the clinical work flow and improve patient outcomes for a particular type of cancer. In addition to precision when predicting the likelihood of experiencing relapses, the model is required to correctly risk stratify the severity of illness for similarly situated patients to prevent unequal treatments when allocating critical healthcare resources. As the patients' main outcome under assessment, *i.e.*, cancer recurrence which is the class label, could be unknown for a portion of the study group (phenomenon known as censorship, cf. a detailed discussion below), existing fairness notions and algorithms which assume the availability of class label become inapplicable.

In this example, we see that the presence of censorship nullifies any bias quantification and mitigation of the class label dependent fair training procedure when the model is deployed. Such censorship can arise in various ways as shown in Figure (1): the individual has not yet experienced the event of interest prior to the study ends so this individual's class label remains unknown, *e.g.*, the individual d_4

is censored; the studied individual is impossible to further follow-up due to various reasons such as withdraw from the study, lost to follow-up during the study period and experience a competing event, *e.g.*, the individual d_2. Moreover, censored data prevails beyond clinical prediction, *e.g.*, marketing analytics (KKBOX dataset) (Kvamme, Borgan, and Scheel 2019), recidivism prediction instrument (COMPAS (Larson and Kirchner 2016) and ROSSI (Fox, Carvalho et al. 2012) dataset), to name a few in which the event of interest/class label can be unknown due to the same reasons discussed in the clinical prediction task (Kvamme, Borgan, and Scheel 2019). However, existing fairness notions and algorithms typically focus on the "processed" benchmark datasets that either drop observations with uncertain class labels due to censorship (Chouldechova 2017; Quy et al. 2022; Hort et al. 2022) or omit the censorship information of these instances (Wan et al. 2020; Vasudevan and Kenthapadi 2020; Zhang and Ntoutsi 2019) which does not center realistic data characteristics, thus preventing "fair" models developed in the lab being applicable in real-world applications. In addition, although individual fairness, compared with group fairness, enjoys the merits of free of sensitive attribute specification and harder to fail by scrutinizing at the finer granular individual level (Barocas, Hardt, and Narayanan 2017), it demands the distance calibration resulted from the Lipschitz condition (Lahoti, Gummadi, and Weikum 2019b). In practice, however, even though metrics evaluating both input and output space similarities can be properly defined by domain experts, such a Lipschitz condition is non-trivial to be specified and has therefore been another major obstacle for wider adoption of existing individual fairness in real-world applications.

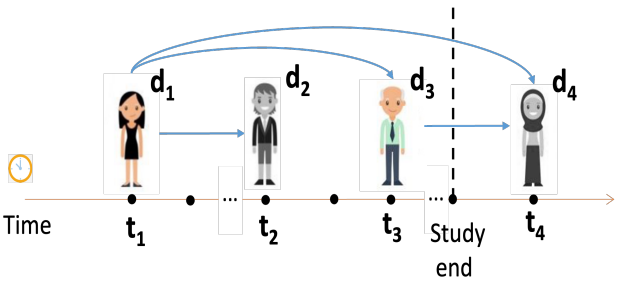

Figure 1: An illustrative example of censoring phenomenon: individuals in grey, *i.e.*, d_2 and d_4, are censored while others, *i.e.*, d_1 and d_3, are non-censored; individuals are arranged in the increasing time order of their survival times with the lowest, $i.e$, t_1, being at left most; the study ends at the time shown as the vertical dash line; there is no edge originate from a censored individual due to censorship.

Summary of our contributions. To tackle the aforementioned challenges, we introduce a new *individual fairness with censorship* setting with the goal of encouraging wider use through attending to the unaddressed challenges of realistic fair model deployment. Explicitly: (i.) We formulate a new research problem for fairness guarantee which relies on the more in line with the realistic assumption that individual outcomes are possibly censored. (ii.) We present a new definition along with a debiasing algorithm which are independent from Lipschitz condition but also are capable of quantifying and mitigating individual unfairness amidst censorship for real-world socially sensitive applications. (iii.) Empirical evaluations on three complete rather than "processed" benchmark datasets confirm the utility of the proposed approach in practice.

Problem Formulation and Related Work

Censored Data

To represent the essence that the data is censored because each individual may eventually experience the event of interest but such information is not present, the censored data can be typically described by three pieces of information: (i.) the observed covariates/features x, (ii.) the survival time T and (iii.) the event indicator δ. The first piece of information characterizes the certain information that is observed for each individual while the possible uncertainty arises from the last two: when δ equals to 1 the event is observed indicating certainty on the event time T or class label (*i.e.*, the event is observed at time T), otherwise the event time is censored resulting unavailability of the class label (*i.e.*, the individual is censored at time T).

AI Fairness

Much progress has been made to quantify and mitigate unfair or discriminatory manner of AI algorithm. These efforts, at the highest level, can be typically divided into two families: *individual fairness* and *group fairness*. A vast majority of existing works focus on group notions, aiming to ensure members of different groups, *e.g.,* gender or race *aka* sensitive attributes, achieve approximate parity of some statistic over class labels, such as statistical parity (Zhang and Ntoutsi 2019), disparate impact (Zafar et al. 2017), equality of opportunity (Hardt et al. 2016) and calibration (Kleinberg, Mullainathan, and Raghavan 2016). Then, related group based debiasing algorithms are designed to enforce respective fairness notions, typically as a constraint or regularizer, and therefore require the availability of class label as well. While enjoying the merit of operational ease, group-based fairness approaches are prone to fail when guaranteeing fairness at the individual level in addition to several other drawbacks (Barocas, Hardt, and Narayanan 2017).

On the other hand, individual fairness alleviates such a drawback through "awareness", requiring individuals who are similarly situated, with respect to the task at hand, receive similar probability distributions over class labels (Dwork et al. 2012). Formally, this objective can be formulated as the Lipschitz property and fairness is achieved iff:

$$\hat{D}(f(x_a), f(x_b)) \leq LD(x_a, x_b) \tag{1}$$

where L is the Lipschitz constant, $D(\cdot)$ and $\hat{D}(\cdot)$ are corresponding functions used to measure the similarly situated in input space, *e.g.,* features x, and similar probability distributions over class labels in output space, *e.g.,* outcomes of the

prediction function $f(\cdot)$, respectively. One of the major obstacles for wider adoption of individual fairness, though, is the distance calibration between the input and output space resulted from L. In addition, existing studies of individual fairness assume availability of the class label, which is impractical in many real-world applications due to the prevailing censorship. Our new fair methodology is a member of this group of individual-based approaches, but resolves these two main questions left open in current literature, thus providing a fairness guarantee across individuals with censorship and the "awareness" is free from Lipschitz condition.

Survival Analysis

The prevalent censored data in real-world AI applications makes survival analysis necessary (Clark et al. 2003; Wang et al. 2021; Turner et al. 2022). For example, to build a model used to aid in the prognosis of disease relapse, the collected raw data will include individuals who experience relapses but also those whose relapse status remains unknown, *i.e.*, censored. A related function commonly used is the *hazard function*, modeling the rate of event occurrence at a specified time t conditioned on surviving to t:

$$h(t|x) = \lim_{\triangle t \to 0} \frac{Pr(t < T < t + \triangle t | T \geq t, x)}{\triangle t} \tag{2}$$

Among the various proposed survival analysis methods, the Cox proportional hazards model (CPH) (Cox 1972) has become the standard for modeling censored data in which the multiplicative relationship between the risk, as expressed by the hazard function and covariates is described, *i.e.*,

$$h(t|x) = h_0(t) \exp(\beta^T x) \tag{3}$$

where $h_0(t)$ is called the baseline hazard function (*i.e.*, when $x = 0$) while β is a set of unknown parameters, which can be estimated by applying the partial likelihood estimation written as follows:

$$L(\beta) = \prod_{T_i \ uncensored} \frac{\exp(\beta^T x_i)}{\sum_{T_j \geq T_i} \exp(\beta^T x_j)} \tag{4}$$

Various approaches have been proposed to model the hazard function from the prevailing censored data (Katzman et al. 2018; Ishwaran et al. 2008; Wang, Li, and Reddy 2019; Bou-Hamad et al. 2011). In addition, care must be taken to ensure the fairness of survival models, the same as other AI approaches. Our work situates in this under-explored research direction to tackle fairness in the presence of censorship. Starting with (Zhang and Weiss 2021, 2022, 2023), there is a different line of work studying fairness with censorship but subject to group-based constraints. Relevantly, the survival model is modified to ensure fair risk predictions as in (Keya et al. 2021). However, their work necessitates the Lipschitz condition as in the conventional individual fairness definitions and does not explicitly considers survival information to address discrimination in the presence of censorship. Our method aims to alleviate these two limitations.

Censored Individual Fairness

To fill the gap between "fair" models in the lab and their deployment in the real-world, this section introduces a first of its kind individual fairness notion that specifically account for censoring while jointly evaluating bias from a ranking perspective to remove the dependence on Lipschitz condition, along with a debiasing algorithm in the presence of censorship.

Quantifying Censored Individual Unfairness

The existing individual fairness notions necessitate the availability of class label while ignoring the censoring information. However, the survival information is important and requires special attention, otherwise substantial bias could be introduced if it is simply neglected. In addition, current individual unfairness quantification depends on the Lipschitz condition for fairness formulation which is non-trivial due to the similarity metric difference between the input and output space. To overcome these, we propose to evaluate unfairness from a ranking perspective while jointly considering survival information for the evaluation of individual fairness with censorship.

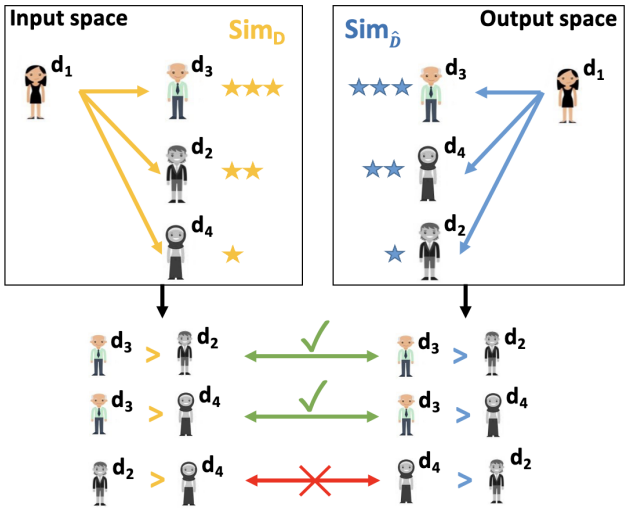

Figure 2: An illustration of quantifying and mitigating individual unfairness from a ranking perspective; Sim_D and $Sim_{\hat{D}}$ are similarity matrices obtained from the input and output space respectively; the number of star(s) next to each individual represents corresponding pairwise level of similarity; a check mark indicates the ranking order is consistent between input and output space while a cross mark means inconsistency.

Discounted Cumulative Fairness. As previously discussed in Example 1, similarly situated patients should receive similar treatment in regard to allocation of critical healthcare resources. Such a fair clinical prediction could be reflected as the ranking order when receiving the service, thus requiring the ranking order in the input space is preserved in the output space as decided by the AI model, which also aligns with the typical individual fairness idea that sim-

ilar individuals situate similarly. Motivated by this, we propose *Discounted Cumulative Fairness* (DCF) to quantify censored individual unfairness from a ranking perspective to alleviate the challenge of Lipschitz condition-based distance calibration. Specifically, DCF first obtains two ranking lists based on the similarity matrices Sim_D and $Sim_{\hat{D}}$ from the input and output space respectively then evaluates the consistency between these two lists. To this end, DCF looks at one individual at a time checking whether other individuals' relative orders to this focused individual are consistent across the input and output space. Take Figure 2 as the example, assume the list derived from Sim_D when focusing on d_1 is $\{d_3, d_2, d_4\}$ (*i.e.*, d_3, d_2, d_4 are the most, second and least similar to d_1) while the encoded list from $Sim_{\hat{D}}$ is $\{d_3, d_4, d_2\}$. As the expected encoded list from $Sim_{\hat{D}}$ should be $\{d_3, d_2, d_4\}$ as well, d_2 is individually unfair treated as d_2 should be ranked closer to d_1 than d_4. Armed with this idea also motivated by learning to rank (Burges, Ragno, and Le 2006), DCF@k is formally defined as below to quantify such a type of inconsistent ranking showing individual unfairness amidst censorship,

$$\text{DCF@k} = \frac{1}{N}\sum_{n=1}^{N}\frac{DCG_{Sim_D(d_n)}(\hat{p})}{DCG_{Sim_D(d_n)}(p)} \tag{5}$$

where N represents the total number of individuals, p and $\hat{p}$ denote the ranking orders based on the input and ouput space simiarilities respectively, and k is the length of the top-k ranking list (we focus on the top-k individuals following the basic principle of individual fairness which only requires similar people are treated similarly), while the formulation for $DCG_{Sim(d_n)}$, which represents *Discounted Cumulative Gain* for each focused individual d_n, is expressed as follows,

$$\underset{\{p,\hat{p}\}\in pos_{Sim_D(d_n)}}{DCG}(pos) = \sum_{pos=1}^{k}\frac{Sim_D(pos)}{ln(pos+1)} \tag{6}$$

where pos is the position of each individual in the ranking list derived from the corresponding similarity matrix for individual d_n characterized by feature x_n, while $Sim_D(pos)$ is the input space similarity between the individual in this position of the ranking list (obtained from either input or output space) and the individual d_n (we will use x_n and d_n interchangeably in the following for ease of expression). Note that the input space similarity matrix Sim_D is often a given apriori as it is problem-specific (Lahoti, Gummadi, and Weikum 2019a,b), while we define $Sim_{\hat{D}}$ as follows,

$$Sim_{\hat{D}}(x_i, x_j) = \frac{1}{1+(1-C_{\triangle}(x_i,x_j))|\bar{h}(t|x_i)-\bar{h}(t|x_j)|} \tag{7}$$

where $\bar{h}(t|x)$ is the hazard function with the base function $h_0(t)$ dropped as it is not individual specific, *i.e.*, $\bar{h}(t|x) = \exp(\beta^T x)$, and $C_{\triangle}(x_i, x_j)$ measures the concordance difference to adjust the similarity between two individuals as evaluated by the hazard function while taking important survival information into consideration.

To understand this key component $C_{\triangle}(x_i, x_j)$, let's consider evaluation amidst censorship as a ranking problem. Specifically, when pairwise comparing one individual x_g with other individuals, *e.g.*, x'_g, the individual with a shorter non-censored time, *i.e.*, $\delta_{t_{\ltimes}} = 1$ in Equation (9), should be assigned a higher hazard score than another individual with a longer survival time, regardless of the longer survival time's censorship status. This can be visualized by means of an order graph as shown in Figure 1 in which edges are originated from individuals with a shorter time and are not censored exclusively, *i.e.*, individual d_1 and d_3, thus reflecting the comparability of pairwise comparisons. In addition, the model should assign a higher hazard score for individual d_1 than all other individuals as well as a higher hazard score for individual d_3 than d_4. This can be interpreted as the fraction of all pairs whose predicted outcome are correctly ordered among all individuals that can actually be ordered. From individual fairness' perspective, similar individuals should receive similar fractions as similar outputs are expected for them from the model, resonating the typical individual fairness idea that similar individuals situated similarly. Ensembling these ideas, $C_{\triangle}(x_i, x_j)$ measuring the concordance difference between x_i and x_j within the corresponding ranking list is mathematically represented as:

$$C_{\triangle}(x_i, x_j) = C_{x_i} - C_{x_j} \tag{8}$$

where the concordance C_{x_n} of individual d_n is defined as:

$$C_{x_n} = \frac{1}{M}\sum_{n'=1}^{k}\mathbb{1}[\bar{h}(t_{\rtimes}|x_{\rtimes}) < \bar{h}(t_{\ltimes}|x_{\ltimes})|\delta_{t_{\ltimes}} = 1] \tag{9}$$

where k is the number of individuals in the ranking list, M is the number of *permissible pair* whose shorter survival time is observed, *i.e.*, M= $\sum \mathbb{1}[\delta_{t_{\ltimes}} = 1]$, and $x_{\rtimes}$ or $x_{\ltimes}$ is the individual with a longer, *i.e.*, $t_{\rtimes} = max(t_n, t'_n)$, or shorter, *i.e.*, $t_{\ltimes} = min(t_n, t'_n)$, survival time.

The concordance difference effectively adjusts the similarity values defined in formula (7). In the general case, we would like the original similarity in the output space to be downscaled according to the prediction deviation as reflected by the concordance difference, which also explicitly includes survival information when quantifying unfairness in the censoring setting.

Align with the existing individual fairness notions, the values of DCF@k is also within the interval of [0,1]. In addition, the higher the DCF@k score, the more consistency between the ranking list encoded from the input and output space and thus, the fairer the model.

Mitigating Censored Individual Unfairness

Our bias mitigation algorithm, *Individual Fair Survival (IFS)* is built upon the deep neural network based survival model *DeepSurv* (Katzman et al. 2018), one of the most popular survival learners. DeepSurv enjoys its popularity as it has sound guarantees of performance when the number of features and interactions increases, in which the linear proportional hazards condition is relaxed to better encode the nonlinearity of censored data. Mathematically, the loss function of DeepSurv is formulated as the negative log partial likelihood:

$$\mathcal{L}(\theta) = -\sum_{i:E_i=1}\left(\bar{h}_\theta(x_i) - \log\sum_{j\in\mathcal{R}(T_i)} e^{\bar{h}_\theta(x_j)}\right) \quad (10)$$

where $N_{E=1}$ is the number of individuals with observed events, $\mathcal{R}(T_i)$ stands for all the remaining patients at time T_i, $\bar{h}_\theta(x_i)$ is the risk output of the neural network and θ are weights of the model.

Such a loss function aims to optimize for predictive performance and does not take fairness into consideration. Here, we propose IFS to generate tailored forecasts while providing fair risk predictions to ensure similar individuals, in the presence of censorship, are treated similarly for real-world socially sensitive applications.

To this end, we devise an individual fairness loss function to mitigate individual unfairness amidst censorship. Specifically, we cast individual fairness guarantee as a ranking problem to alleviate the existing individual fairness approaches' drawback on the dependence of non-trivial Lipschitz condition, while jointly dealing with the censoring of the data. This in practice can be instantiated as the constraint enforcing consistency between the two ranking lists that are obtained from the input and output space, respectively. Still in Figure 2, the first two relative ranking order, *i.e.*, d_3 vs d_2 and d_3 vs d_4, are consistent in the input and output space, while d_2 and d_4 swap their respective order in the input and output space, *i.e.*, inconsistent. The loss function should thus promote the first two relative ranking order while penalizing the last one. Motivated by this, we formulate the loss function as a probabilistic function below,

$$\mathcal{F}(k) = \sum_{n=1}^{N}\sum_{i,j}\mathcal{F}_{d_i,d_j}(d_n) \quad (11)$$

where N is the number of individuals, d_i and d_j are two among the top ranked k individuals from the input space focusing on each individual d_n, and $\mathcal{F}_{d_i,d_j}(d_n)$ is the cross-entropy loss on order consistency based probability distribution difference from the input and output space:

$$\mathcal{F}_{d_i,d_j}(d_n) = -P_{d_i,d_j}log\hat{P}_{d_i,d_j} - (1-P_{d_i,d_j})log(1-\hat{P}_{d_i,d_j}) \quad (12)$$

where P_{d_i,d_j} and $\hat{P}_{d_i,d_j}$ are the probability scores of the relative order in the input and output space:

$$P_{d_i,d_j} = \begin{cases} 1, & Sim_D(d_n,d_i) > Sim_D(d_n,d_j) \\ 0.5, & Sim_D(d_n,d_i) = Sim_D(d_n,d_j) \\ 0, & Sim_D(d_n,d_i) < Sim_D(d_n,d_j) \end{cases} \quad (13)$$

$$\hat{P}_{d_i,d_j} = \frac{1}{1+e^{-(Sim_{\hat{D}}(d_n,d_i) - Sim_{\hat{D}}(d_n,d_j))}} \quad (14)$$

Intuitively, P_{d_i,d_j} formulates the known probability, typically given as a priori, on whether d_i is more similar to d_n than d_j when centering on d_n in the input space, while $\hat{P}_{d_i,d_j}$ does the same but in the output space based on model prediction based probability. Then, this similarity based relative ranking difference in the input and output space represents the ranking inconsistency loss and is quantifies as $\mathcal{F}_{d_i,d_j}(d_n)$. Last, $\mathcal{F}(k)$ aggregates the loss over all individuals. As a note, the proposed $\mathcal{F}_{d_i,d_j}(d_n)$ encourages the consistency between the two ranking lists that are obtained from the input and output space, respectively. Enforcing this, individuals' similarity in the input space (e.g., severity of the illness) will be preserved in the output space (e.g., allocating critical healthcare resources), thus encouraging similar individuals (in the input space) being treated similarly (in the output space) while accounting for censorship. In addition, IFS focuses on the top-k ranking to encourage locally similar without requiring global similarity as individual fairness only asks for similar outcomes for similar individuals.

With these, the overall objective function of IFS, to be minimized, can be formulated as:

$$\mathcal{L}(\theta, k) = \mathcal{L}(\theta) + \lambda\mathcal{F}(k) \quad (15)$$

where λ is the trade-off parameter controlling the strength of individual fairness constraint.

Charac. \ Dataset	ROSSI	COMPAS	KKBOX
Sample #	432	10,325	2.8M
Censored%	0.736	0.732	0.347
Feature #	9	14	18

Table 1: The summary of datasets for empirical evaluations.

Empirical Evaluations

Datasets

We evaluate our approach on three real-world datasets explicitly include survival information and with socially sensitive concerns: i) The *ROSSI* datase (Fox, Carvalho et al. 2012) comprises information on 432 convicts who were discharged from a Maryland state prison during the 1970s and monitored for a year following their release. The study involved a randomized experiment in which half of the randomly assigned individuals were given financial assistance, while the other half received no aid. Roughly 73.6% of the dataset comprises censored observations, all of which are censored at the 52-week mark. ii) The *COMPAS* dataset (Larson and Kirchner 2016), a landmark dataset in algorithmic unfairness, bears similarities to the ROSSI dataset used to forecast recidivism of convicts released from Broward County, Florida, but with significantly larger sample size of 10,325 instances. Additionally, like the ROSSI dataset, the COMPAS dataset exhibits a censored rate of 73%. iii) The *KKBOX* dataset from the WSDM-KKBox's Churn Prediction Challenge 2017 (Kvamme, Borgan, and Scheel 2019). Its objective is to predict whether a user of KKBox, which is a music streaming service, will renew their subscription within 30 days after the expiration of their current streaming subscription. With a relatively low censored

Datasets	Method \ Metrics	DCF@10%↑	C-index% ↑	Brier Score%↓	Time-dependent AUC%↑
ROSSI	FDCPH	*44.12*	55.81	19.83	76.18
	CPH	33.41	64.24	17.67	77.12
	RSF	36.17	65.56	15.12	**79.32**
	DeepSurv	31.43	**66.67**	**14.71**	*78.18*
	IFS	**57.68**	*65.87*	*15.05*	77.85
COMPAS	FDCPH	72.27	63.54	24.12	65.16
	CPH	73.51	69.24	20.35	65.15
	RSF	*74.64*	72.61	15.62	71.76
	DeepSurv	74.18	**75.12**	**13.42**	**71.83**
	IFS	**86.78**	*73.17*	*13.89*	*71.77*
KKBOX	FDCPH	*58.64*	70.44	21.23	69.73
	CPH	47.32	80.02	18.17	72.95
	RSF	42.41	82.32	**14.24**	78.18
	DeepSurv	43.45	*83.01*	*14.33*	*80.71*
	IFS	**69.64**	**84.15**	14.35	**81.12**

Table 2: Comparison of IFS against various baselines on diverse datasets with best results marked in bold and second best in *italics*. A higher value is desired for the metrics followed by an ↑ while followed by ↓ are the opposite.

rate of 34.7%, the KKBox dataset comprises over 2.8 million subscribers. Table 1 summarizes their statistics and properties.

IFS Implementation

The IFS is optimized with Adam optimizer via backpropagation and automatic differentiation, with learning rate 0.01 using PyTorch, and in a mini-batch setting for 50 epochs with a mini-batch size of 128. In addition, the number of top k in the ranking list is set as 10 while λ as 1 in the overall objective function for quantitative performance comparison. IFS follows the hyperparameter settings (e.g., hidden unit number) of our base model DeepSurv (Katzman et al. 2018) and further does grid search for fairness specific tuning parameters (the search space of k is 4-50 and λ is $1e^{-4}$-$1e^{4}$).

Benchmark Performance

This section first investigates the theoretical design of IFS. For comparison: i) we implemented the recently proposed fair survival model *FDCPH* (Keya et al. 2021) which is the only work touching on debiasing across censored individuals, to the best of our knowledge (note that only the most competitive one is considered among different variants proposed therein). We also compare against: ii) the commonly used survival analysis model *CPH* (Cox 1972), iii) the state of the art random survival forests (*RSF*) (Ishwaran et al. 2008) which is a meta-estimator that builds multiple survival trees on different subsets of a given censored dataset, as well as: iv) deep neural network based *DeepSurv* to encode the nonlinearity of censored data (Katzman et al. 2018) which is also the base model of our method as the additional baselines. Other competing fairness methods are not considered as none of them can be transferred to censored settings. Neither are group-based fair survival models (Larson and Kirchner 2016; Zhang and Weiss 2021, 2022; Sonabend et al. 2022) as they necessitate the extra effort in specifying sensitive attribute for bias mitigation which is unspecified in individual fairness setup.

In addition to the proposed censored individual fairness metric, we also report the commonly used survival model utility metrics: i) the *C-inde*x which equals the area under ROC curve (AUC) in the absence of censorship (Harrell et al. 1982); ii) the *Brier score* also attends to the calibration of the model by quantifying the mean squared difference between the predicted probability and the actual outcome (Brier and Allen 1951); and iii) the *Time-dependent AUC* testing the discriminative power of the model when distinguishing individuals who experience the event of interest from those who have not up to the time t (Chambless and Diao 2006). Furthermore, the similarity matrix in the input space Sim_D, as per standard, is given as a priori (Lahoti, Gummadi, and Weikum 2019b,a). To show the generalization of IFS, we construct Sim_D using the euclidean distance with feature scaling (Han, Pei, and Tong 2022). In addition, the number of top k in the ranking list is set as 10 while λ as 1 in the overall objective function for quantitative performance comparison. All methods are trained the same way for fair comparison with the 5-fold cross validation results summarized in Table 2.

As shown in Table 2, our new IFS method dominates all other baseline models in terms of mitigating bias in the presence of censorship, and is second-best on the majority of model utility metrics. We note that IFS' superior discrimination minimizing capability over other second-best baselines can be as high as 73.07%, while still being highly competitive with narrow margins being at most 2.6% within the top utility performer. This shows the desirable fairness-utility trade-off of our approach amidst censorship. On the other hand, the inferior performance of FDCPH shows the drawbacks of Lipschitz condition based distance calibration as

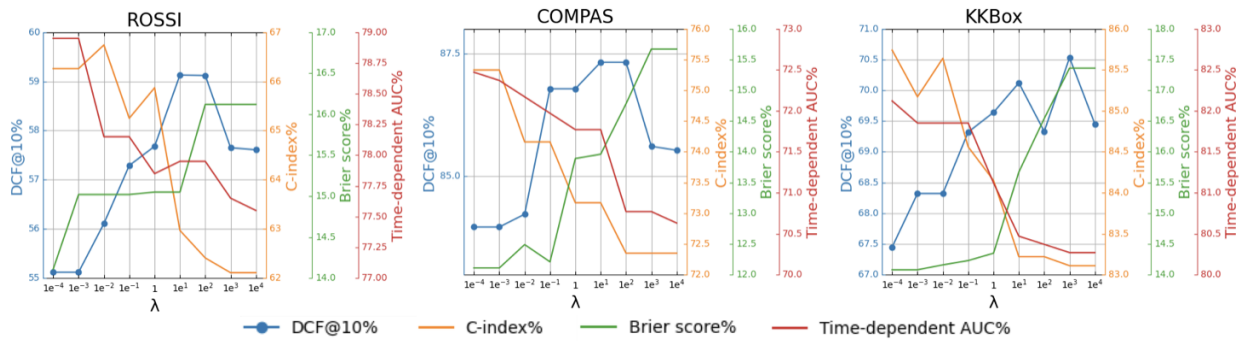

Figure 3: The model utility and individual fairness trade-off fined grained by the tunable parameter λ.

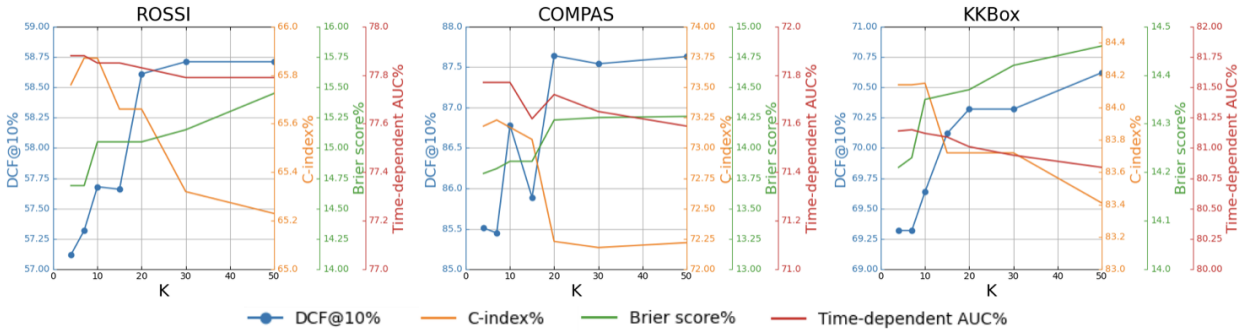

Figure 4: The effects on the choice of k on model utility and individual fairness.

a result of variation in data as well as not explicitly considering survival information in the model design. This also demonstrates that fairness amidst censorship cannot be trivially solved by a simple combination of existing fairness approaches in the absence of censorship techniques.

The Effect of λ and K on Model Utility and Individual Fairness

The design of IFS also provides a clear mechanism to fine-tune the trade-off between utility and fairness, allowing the end-user to adjust the model when the initial model does not meet the discrimination or utility requirements. To illustrate this mechanism, Figure 3 shows the results of adjusting the value of λ, which controls the degree of trade-off between utility and fairness in the IFS model. As we can see, when λ is set as a small value, the model utility and individual fairness performance are not significantly impacted. This suggests that the model is able to achieve an excellent balance between utility and individual fairness when λ is small. When the value of λ further increases, the model's performance in promoting individual fairness is on the rise then reaches a peak or drops while the utility declines. This could be due to the fact that top-ranked individuals are hard to be obtained within reasonable training epochs when λ is relatively large. Clients can therefore explore an appropriate λ to accommodate their realistic needs according to their respective constraints.

We also investigate the effect of different top-k values as shown in Figure 4. From the obtained results, a larger value of k typically leads to increased individual fairness performance while model utility remains relatively steady. These results are expected as the identification of top-k is mainly for optimizing the individual fairness part of the overall learning function. This also validates the previous results on benchmark performance and the effect of λ demonstrating that IFS achieves a desirable fairness-utility trade-off amidst censorship. It can be noted that, compared to other metrics, the DCF performance of IFS shows more frequent fluctuations during the parameter variation process. This indicates the need for an additional fairness metric that takes the stability perspective into account, in order to comprehensively evaluate the model's performance.

Conclusion

Given the observed gap between the prevailing real-world applications with censorship and the assumption of class label availability of existing AI fairness methods, this work made an initial investigation on individual fairness with censorship. In addition, this work also took a step further to quantify and mitigate individual unfairness from a ranking perspective, thus alleviating the drawback of the non-trivial Lipschitz constant specification of the existing individual fairness studies. The proposed notion and algorithm are expected to be versatile in quantifying and mitigating bias in various real-world socially sensitive applications. In addition, this work defines a new task and opens possibilities for future work on a comprehensive study of AI fairness.

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Woodruff, A.; Luu, C.; Kreitmann, P.; Bischof, J.; and Chi, E. H. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. *AIES*.

Bou-Hamad, I.; Larocque, D.; Ben-Ameur, H.; et al. 2011. A review of survival trees. *Statistics surveys*, 5: 44–71.

Brier, G. W.; and Allen, R. A. 1951. Verification of weather forecasts. In *Compendium of meteorology*, 841–848. Springer.

Burges, C.; Ragno, R.; and Le, Q. 2006. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19.

Chambless, L. E.; and Diao, G. 2006. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in medicine*, 25(20): 3474–3486.

Chang, V. 2021. An ethical framework for big data and smart cities. *Technological Forecasting and Social Change*, 165: 120559.

Chen, I. Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; and Ghassemi, M. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, 4.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Clark, T. G.; Bradburn, M. J.; Love, S. B.; and Altman, D. G. 2003. Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, 89(2): 232–238.

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

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Fox, J.; Carvalho, M. S.; et al. 2012. The RcmdrPlugin. survival package: Extending the R Commander interface to survival analysis. *Journal of Statistical Software*, 49(7): 1–32.

Han, J.; Pei, J.; and Tong, H. 2022. *Data mining: concepts and techniques*. Morgan kaufmann.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *Jama*, 247(18): 2543–2546.

Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.

Hort, M.; Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2022. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; Lauer, M. S.; et al. 2008. Random survival forests. *Annals of Applied Statistics*, 2(3): 841–860.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 1–12.

Keya, K. N.; Islam, R.; Pan, S.; Stockwell, I.; and Foulds, J. 2021. Equitable allocation of healthcare resources with fair survival models. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 190–198. SIAM.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kvamme, H.; Borgan, Ø.; and Scheel, I. 2019. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129): 1–30.

Lahoti, P.; Gummadi, K.; and Weikum, G. 2019a. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proceedings of the VLDB Endowment*, 13(4): 506–518.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019b. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, 1334–1345. IEEE.

Larson, S. A. J.; and Kirchner, L. 2016. There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Meyer, D. 2018. Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women. Fortune, October 10.

Miller, C. C. 2015. Can an algorithm hire better than a human. *The New York Times*, 25.

Quy, T. L.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Data Mining and Knowledge Discovery*.

Saxena, N. A.; Zhang, W.; and Shahabi, C. 2023a. Missed Opportunities in Fair AI. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

Saxena, N. A.; Zhang, W.; and Shahabi, C. 2023b. Unveiling and Mitigating Bias in Ride-Hailing Pricing for Equitable Policy Making. *arXiv preprint arXiv:2301.03489*.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Skirpan, M.; and Gorelick, M. 2017. The Authority of "Fair" in Machine Learning. In *FAT ML Workshop*.

Sonabend, R.; Pfisterer, F.; Mishler, A.; Schauer, M.; Burk, L.; and Vollmer, S. 2022. Flexible Group Fairness Metrics for Survival Analysis. *arXiv preprint arXiv:2206.03256*.

Turner, K.; Brownstein, N. C.; Thompson, Z.; El Naqa, I.; Luo, Y.; Jim, H. S.; Rollison, D. E.; Howard, R.; Zeng, D.; Rosenberg, S. A.; et al. 2022. Longitudinal patient-reported outcomes and survival among early-stage non-small cell lung cancer patients receiving stereotactic body radiotherapy. *Radiotherapy and Oncology*, 167: 116–121.

Vasudevan, S.; and Kenthapadi, K. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2773–2780.

Wan, C.; Chang, W.; Zhao, T.; Cao, S.; and Zhang, C. 2020. Denoising Individual Bias for Fairer Binary Submatrix Detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2245–2248.

Wang, P.; Li, Y.; and Reddy, C. K. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6): 1–36.

Wang, X.; Zhang, W.; Jadhav, A.; and Weiss, J. 2021. Harmonic-Mean Cox Models: A Ruler for Equal Attention to Risk. In *Survival Prediction-Algorithms, Challenges and Applications*, 171–183. PMLR.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zhang, W.; Bifet, A.; Zhang, X.; Weiss, J. C.; and Nejdl, W. 2021. FARF: A Fair and Adaptive Random Forests Classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 245–256. Springer.

Zhang, W.; and Ntoutsi, E. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1480–1486.

Zhang, W.; and Weiss, J. 2021. Fair Decision-making Under Uncertainty. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE.

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

Zhang, W.; and Weiss, J. C. 2023. Fairness with censorship and group constraints. *Knowledge and Information Systems*, 1–24.

Žliobaitė, I. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4): 1060–1089.