ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230621

Individual Fairness Under Uncertainty

Wenbin Zhang[a;*], **Zichong Wang**[a], **Juyong Kim**[b], **Cheng Cheng**[b], **Thomas Oommen**[c], **Pradeep Ravikumar**[b] **and Jeremy Weiss**[d]

[a]Florida International University, Miami, FL 33199
[b]Carnegie Mellon University, Pittsburgh, PA 15213
[c]University of Mississippi, Oxford, MS, 38677
[d]National Institutes of Health, Bethesda, MD 20892

Abstract. Algorithmic fairness, the research field of making machine learning (ML) algorithms fair, is an established area in ML. As ML technologies expand their application domains, including ones with high societal impact, it becomes essential to take fairness into consideration during the building of ML systems. Yet, despite its wide range of socially sensitive applications, most work treats the issue of algorithmic bias as an intrinsic property of supervised learning, *i.e.*, the class label is given as a precondition. Unlike prior studies in fairness, we propose an individual fairness measure and a corresponding algorithm that deal with the challenges of uncertainty arising from censorship in class labels, while enforcing similar individuals to be treated similarly from a ranking perspective, free of the Lipschitz condition in the conventional individual fairness definition. We argue that this perspective represents a more realistic model of fairness research for real-world application deployment and show how learning with such a relaxed precondition draws new insights that better explains algorithmic fairness. We conducted experiments on four real-world datasets to evaluate our proposed method compared to other fairness models, demonstrating its superiority in minimizing discrimination while maintaining predictive performance with uncertainty present.

1 Introduction

There is recent concern that we are in the midst of a discrimination crisis within the field of machine learning (ML) and artificial intelligence (AI) [3, 38, 40]. Rightfully, the AI/ML community has conducted vast research to study the quantification and mitigation of algorithmic bias, which is critical for the use of algorithmic decision-making systems in domains of high societal impact such as criminal justice [10], healthcare [9], predictive policing [39], and employment [40]. Thus far, most studies tackle the problem by proposing fairness constraints via regularizers/optimizations at the group level: first identify a *sensitive attribute*, *e.g.*, race or gender, as a potential source of bias among the collection of high-level groups; then achieve parity for some fairness statistics of the classifier, *e.g.* the prediction accuracy and true positive rate, across the predefined groups [30]. These group fairness approaches, however, are inapplicable when class label uncertainty is present [44]. Additionally, while group fairness enjoys the merit of easy operationalization, its aggregative characteristic makes it easy to fail [2].

* wenbin.zhang@fiu.edu

In contrast, the *individual fairness* approach alleviates these drawbacks by evaluating a finer granularity of fairness at individual level. The compelling notion of individual fairness is proposed in the seminal work of [15], which requires similarly situated individuals to receive similar probability distributions over class labels to prevent inequitable treatment. Individual fairness, without the need to explicitly identify sensitive attributes, is much less restrictive than group fairness. However, the Lipschitz condition required in existing individual fairness literature is a nontrivial task to satisfy, as 1. the Lipschitz constant specification is hard due to the difference in distance metrics between the input and outcome spaces; 2. distance calibration is required as the absolute distance comparison in the Lipschitz condition tends to fail in calibrating differences among different individuals [13]. Such difficulty was also pointed out in [31] but only resulted in additional efforts of metric learning, whereas our rank-based method removes the need of the Lipschitz constant and distance calibration by avoiding the absolute distance value comparison.

Another major obstacle in the real-world applicability of individual (and also group) fairness is the assumption of full class label availability, which fails when there is uncertainty in class labels due to *censoring*, a phenomenon where the information about the event of interest is partially known [44, 24, 26]. Considering an example from a clinical prediction task (Figure 1), for censored individuals d_2 and d_4, the true time to relapse or hospital discharge is unknown, causing the uncertainty in class labels. Due to the inability to handle censorship information, existing fairness studies quantify and mitigate bias by focusing on the proportion of data with assured class label, thus either dropping observations with uncertain class labels [10, 14, 45] or omitting the censorship information [35, 34, 43]. However, removing them would bias the results towards the individuals with known class labels [33].

In summary, there is a need for an algorithm that addresses individual fairness in ML under uncertainty, an under-explored area of research, with two requirements: **i) Free from the Lipschitz condition resulting from the principle of individual fairness.** Without this, the algorithm may have limited use cases due to the metric calibration between the input and output spaces. **ii) Quantifying and mitigating bias under uncertainty.** The algorithm should not ignore the uncertainty in censored data or the censorship information to avoid bias.

To tackle the aforementioned issues, this paper conducts an initial investigation of *individual fairness under uncertainty* for a fairness guarantee more in line with realistic assumptions across indi-

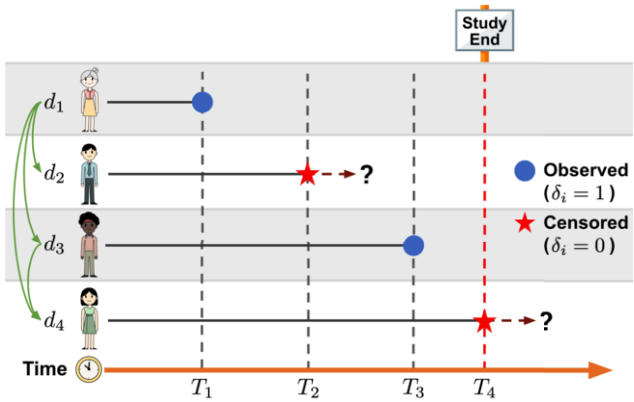

Figure 1: An illustration of the censoring phenomenon. Individuals d_2 and d_4 are censored while others, *i.e.*, d_1 and d_3, are non-censored. Individuals are arranged in the increasing time order of their survival times with the lowest, $i.e$, T_1, being at topmost. The study ends at the time shown as the red vertical dash line. There is no edge originating from a censored individual due to censorship, which means that pair comparison between two individuals cannot be made when the individual with lower survival time is censored.

viduals and free from the Lipschitz condition. Our individual fairness measure, named *Fair Normalized Discounted Cumulative Gain* (*FNDCG*), is motivated by the same individual fairness principle [15] that similar individuals should be treated similarly, while formulated as the correlation of similarities in the feature and risk spaces respectively, establishing a new fairness measure usable on censored data. Along with FNDCG, we also propose a corresponding algorithm to address discrimination involving censored individuals. Our method, named *fairIndvCox*, augments the standard model of survival analysis, the Cox proportional hazard model, by being aware of individual fairness while learning the parameters of risk prediction.

To our knowledge, this work is the first attempt to quantify and mitigate bias under the individual fairness principle, but from a ranking perspective, with uncertainty present, and *free of* the Lipschitz condition. Our major contributions are summarized as follows:

- We formulate a new research problem of individual fairness guarantee in learning with uncertainty.
- We devise *FNDCG*, a new notion of individual fairness to measure bias on censored data. Defined with the correlation of similarity in the feature space and the one in the risk prediction space, FNDCG does not require Lipschitz condition and complete class labels.
- We propose a debiasing algorithm named *fairIndvCox* for bias mitigation in censorship settings, by incorporating our individual fairness measure into the standard model of survival analysis.
- We evaluate our debiasing algorithm on four real-world datasets with censorship, comparing it with four survival analysis algorithms and its Lipschitz variant. This confirms the utility of the proposed approach in practice. Further analysis also illustrates the trade-off between individual fairness and predictive performance.

The remainder of this paper is organized as follows. In Section 2, we describe related work in fair machine learning and learning with uncertainty, followed by the preliminaries of survival analysis and the problem definition in Section 3. In Section 4, we propose our notion of individual fairness under uncertainty and corresponding survival model with an individual fairness specification. In Section 5, we empirically validate the effectiveness of our learning algorithm on real-world survival analysis datasets and provide qualitative analysis on the effect of the hyper-parameters on the model. Finally, we conclude and provide future directions in Section 6.

2 Related Work

2.1 Censored Data

In many real-world applications, the main outcome under assessment, *i.e.*, the class label, could be unknown for a portion of the study group. This phenomenon, deemed censorship, can arise in various ways, hindering the use of many algorithms. For example (Figure 1), a study may end before an individual experiences the event of interest, *e.g.*, individual d_4. The studied individual can also be lost to follow-up during the study period, withdraw from the study, or experience a competing event making further follow-up impossible, *e.g.*, individual d_2. In the typical setting of survival analysis, censored examples are only guaranteed not to have experienced events until their last observation, *e.g.* t_2 and the end of the study for d_2 and d_4, respectively, and we do not know their exact class labels.

The censorship information is used together with the observed data to fit or evaluate survival models, a statistical model that analyzes the expected duration of time until each individual's event. Specifically, we can guarantee that a censored example with the time of event T happens after T, so we can compare two events at T_1 and T_2 for $T_1 < T_2$ if neither is censored at T_1, regardless of censorship at T_2. For instance, the green edges in Figure 1 represent the comparable pairs among individuals with censored and observed events (as the order graph), from which we can tell that d_1 happens before d_2, while whether d_2 happens before d_3 or not remains unknown.

Given that censored data is common, *e.g.*, clinical prediction (Support) [25], marketing analytics (KKBox) [26], recidivism prediction instrument datasets (COMPAS [1] and ROSSI [17]), survival analysis has gained popularity in applied work. For example, in customer analytics whether a customer will cancel the service, *e.g.,* event of interest/class label, can be unknown due to various reasons discussed above [26]. Similarly, one may predict in domains of reoffense [1], analyzing financial outcomes in actuarial analysis [36], and predictive maintenance in mechanical operations [37].

2.2 Fairness in AI

Quantifying Bias Much progress has been made to quantify and mitigate unfair or discriminatory behaviours of AI algorithms. These efforts, at the highest level, can be typically divided into two families: *individual fairness* and *group fairness*. A vast majority of existing works focuses on the group notions, aiming to ensure members of different groups, *e.g.,* gender or race *aka* sensitive attributes, achieve approximate parity of some statistics over class labels, such as statistical parity [43], disparate impact [39], and equality of opportunity [18]. While enjoying the merit of easy operationalization, group-based fairness methods fail at guaranteeing fairness at the individual level in addition to several other drawbacks [2].

Individual fairness, on the other hand, alleviates such a drawback by requiring that individuals who are similarly situated with respect to the task at hand receive similar probability distributions over class labels [15]. Formally, this objective can be formulated as the Lipschitz property, and fairness is thus achieved iff:

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

where L is the Lipschitz constant, $D'(\cdot,\cdot)$ and $D(\cdot,\cdot)$ are corresponding distance functions of features in input space, x, and prob-

ability distributions over class labels in output space, $f(\cdot)$, respectively. The major obstacles for wider adoption of individual fairness, though, are the difficulty of calibrating the distance functions resulted from the Lipschitz condition and the assumption of the availability of class labels, which is impractical in many applications due to censorship. For instance, in the ML-task of predicting critical illness in COVID-19 patients [29], clinical knowledge is required to calibrate the distance-based comparison in Equation 1 since a 10-year difference in age ($D'(\cdot,\cdot)$) for patients younger than 25 would likely result in not much of a difference in risk outcomes ($D(\cdot,\cdot)$), whereas a 10-year difference ($D'(\cdot,\cdot)$) for patients older than 65 could lead to a significant increase in the risk of progressing to critically ill ($D(\cdot,\cdot)$). In addition, a patient may experience censorship, introducing uncertainty about the true progression of their illness at the time of evaluation.

Our new individual fairness methodology resolves these two main limitations in current literature, providing a fairness guarantee across individuals with censorship and is free from the Lipschitz condition.

Mitigating Bias The fairness notions mentioned above are used as a constraint or as a regularizer to enforce fairness. These debiasing algorithms, mostly group-based, can be categorized into three groups based on the stage where machine learning intervention happens: the pre-processing, in-processing, and post-processing groups.

The first group, pre-processing approaches, works on bias in the data or input stage, assuming that unbiased training data is accessible for a fair ML model. These methods modify the data distribution to ensure fairness of the representations from different groups and are model-agnostic. Examples of this group include data massaging [21], which changes data distribution, an extension called local massaging [46], and reweighing [7], which assigns different weights to the communities.

The second group, in-processing approaches, directly changes ML algorithms to produce unbiased predictions and is generally model-specific. For example, in [43], the fairness gain is incorporated into the splitting criteria of the Hoeffding Tree algorithm, which is later extended in [42] to ensemble-based methods. In [45], the Mann Whitney U test is applied to fairness learning in multi-task regression. These methods focus on group fairness and require complete class labels. Yet, there is a limited number of research on individual fairness under data censorship, which this work focuses on.

The last group, post-processing approaches, modifies the decision boundaries to fairly represent diverse groups. Examples include building an interpretable model [41], adjusting the decision threshold to reduce unfairness [18], and moving decision boundaries of the deprived communities to prevent discrimination [16]. However, applying these techniques under censorship is not straightforward, as decision boundaries may also be censored owing to their distribution.

2.3 Survival Analysis

The prevalence of censored data motivates the study of survival analysis to address the problem of partial survival information from the study cohort [11]. The Cox proportional hazard (CPH) model [12] is the most commonly used method, which expresses the hazard function as the product of a shared time-dependent baseline hazard and an individual-specific risk function. Developing the CPH model, [23] parameterized the effect of an individual's covariates by a neural network. Another line of research is tree-based methodology [4, 20], where the splitting rule is modified to handle censored data and is free from the proportional assumption of the CPH model. Interested readers may refer to [36] for a comprehensive survey on recent methods of modeling censored data.

Like other AI approaches, care must be taken to ensure the fairness of survival models to prevent bias against deprived communities. Starting with [44], there is a line of work studying fairness with censorship but subject to group-based constraints. In addition, the survival model is modified to ensure fair risk predictions as in [24]. However, their work requires the Lipschitz condition as in conventional individual fairness and does not explicitly consider the survival information to address discrimination. Our method aims to address these two limitations.

3 Notations and Problem Definition

In this section, we provide preliminary notations and concepts of survival analysis, followed by the definition of the problem of our concern. In survival analysis, censored data can be typically described as follows. Each individual d_i with index $i \in \{1, \cdots, N\}$ is equipped with a characteristics tuple (x_i, T_i, δ_i), where the entries of each tuple are i) x: the observed feature, ii) T: the survival time, *i.e.* the time of event, and iii) δ: the event indicator, which indicates whether the event is observed. In the setting of survival analysis, the event is observed only when $\delta = 1$, and T is the actual time of event. When $\delta = 0$, the event time is censored, resulting in uncertainty on the class label and is only known to be greater than or equal to T.

The modeling function commonly used is the *hazard function*, which specifies the instantaneous rate of event occurrence at a specified time t conditioned on surviving to t:

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

Given a hazard model, one can also compute the survival function $S(t|x) = \Pr(T > t|x)$, the probability that the event occurs after a specific time t by

$$S(t|x) = \exp\left(-\int_0^t h(t|x)dt\right) \tag{3}$$

Among the various proposed survival analysis methods, the Cox proportional hazards model (CPH) [12] has become the standard for modeling censored data, which defines the relation between the hazard function $h(t|x)$ and the covariates as:

$$h(t|x) = h_0(t)\exp(\beta^\top x) \tag{4}$$

where $h_0(t)$ is called the baseline hazard function (*i.e.*, when $x = 0$), and β is a set of unknown parameters which can be estimated by applying the maximum likelihood estimation. Given a dataset of N individuals $\{(x_i, T_i, \delta_i)\}_{i=1}^N$ with i.i.d. assumption, we can compute the likelihood as the product of the likelihood of the uncensored individuals. Such function is called the partial likelihood and can be written as follows:

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta^\top x_i)}{\sum_{j:T_j \geq T_i} \exp(\beta^\top x_j)} \tag{5}$$

The partial likelihood estimate $\hat{\beta} = \arg\max_\beta L(\beta)$ can be obtained by maximizing the partial likelihood function. Note the partial likelihood function does not include the baseline hazard function. One can also add a regularization function, such as ridge or lasso regularization, for β.

To evaluate survival models, the *concordance index*, or *C-index*, is commonly used [19]. Given a survival model, the C-index of the model measures the fraction of all comparable pairs of individuals whose predicted survival times are correctly ordered as training data:

$$C = \frac{1}{\sum_{i:\delta_i=1} |\{j : T_j > T_i\}|} \sum_{i:\delta_i=1} \sum_{j:T_j>T_i} \mathbb{1}[f(x_j) > f(x_i)] \quad (6)$$

where $f(x)$ is the expected survival time for an individual [32]. C-index is also equal to the area under the ROC curve (AUC) in the presence of censorship. In a proportional hazard model, the order of expected survival time is the same as the order of the hazard function. Please see Figure 1 for an example of an order graph, which represents the comparable pairs of individuals.

The main problem we address in this work is to devise an algorithm that can quantify the individual fairness notion in survival analysis and use the quantification to mitigate the bias. Under the general assumption of survival analysis, unlike most existing works of individual fairness, not all individuals are given a label, or survival time, due to data censoring. Another desirable quality the algorithm has is to alleviate or be free from the Lipschitz condition, enabled by the locality between similarity metrics. Note that although similarity-based constraints have been formulated to alleviate bias [22, 13], we are the first to make the contribution of taking censored information into consideration while establishing our similarity-based constraint.

4 Method

We introduce a learning algorithm for individual fairness with censored data. In Section 4.1, we define a rank-based similarity measure of risk scores and propose a corresponding individual fairness score, named *FNDCG@k*. In Section 4.2, we propose a survival analysis model, named *fairIndvCox*, which incorporates FNDCG@k into the Cox proportional hazard model.

4.1 Individual Fairness Notion under Uncertainty

Existing individual fairness notions depend on the Lipschitz condition, which is non-trivial due to the difference in the similarity metrics of the input and output spaces. In addition, they do not consider survival information when quantifying unfairness, which is important and requires special attention; otherwise, substantial bias could be introduced. To overcome these, we propose to evaluate unfairness from a ranking perspective while jointly considering survival information.

For each individual, we obtain two ranked lists of other individuals based on the similarity matrix $\mathrm{Sim}_{D'}$ (on the input space) and Sim_D (on the output space), and require the relative *order* of individuals in these two lists to be consistent with each other. The intuition still follows conventional individual fairness as similar individuals should have similar prediction results, but approaching it from a ranking perspective instead of the absolute distance value comparison (Equation 1) promotes applicability by avoiding Lipschitz specification and distance calibration. For instance (Figure 1), assume the ordered list derived from $\mathrm{Sim}_{D'}$ between patient d_1 and three other patients is $\{d_3, d_2, d_4\}$, ordered by closest-to-farthest. Then, the predictions are individually fair for d_1 if the encoded list from Sim_D is $\{d_3, d_2, d_4\}$ as well, *i.e.* fairness is obtained when patients, ordered by their similarity to patient d_1, have predicted risks in the same order of similarity to d_1's risk. This potentially results in a patient more similar to d_1 receiving a more similar treatment as d_1. Note that the input similarity matrix $\mathrm{Sim}_{D'}$ is often given a priori as it is problem-specific [28, 27], while we define Sim_D as follows,

$$\begin{aligned}\mathrm{Sim}_{D,ij} &= \mathrm{Sim}_D(d_i, d_j) = \exp\left(-|\bar{h}(t|x_i) - \bar{h}(t|x_j)|\right) \\ &= \exp\left(-|\exp(\beta^\top x_i) - \exp(\beta^\top x_j)|\right) \end{aligned} \quad (7)$$

$\mathrm{Sim}_{D,ij}$ is the (i,j)-th entry of Sim_D and $\bar{h}(t|x)$ is the hazard function with $h_0(t)$ dropped, *i.e.*, $\bar{h}(t|x) = \exp(\beta^\top x)$, as it is not individually specific in the CPH model.

In Equation 7, the similarity metric is formulated as the exponential of the negative difference of the risk score. We make a note that this considers various factors to make a similarity metric that performs a trade-off between accuracy and fairness. First, the exponential followed by negation is used for smoothing. This bounds the difference in the unbounded risk scores to a value between 0 and 1. Second, it transforms a distance metric into a similarity function, which has a value closer to 1 when the two individuals are similar. It also makes the function applicable to discounted cumulative gain (DCG) [6], which will be used to compute the fairness quantification. In DCG@k, the quality of the most similar pairs in the output space will be accumulated with a discounted factor decaying with their ranking. Here, similarity is more proper than a metric for the quality function as the closer a pair is, the higher the function is.

Since the encoded ranking list should also take important survival information and consistency between predicted and actual outcome into consideration, we adjust Sim_D according to the concordance difference ($C_\triangle$):

$$\mathrm{Sim}_{D,ij} = (1 - C_\triangle(x_i, x_j)) \exp\left(-|\exp(\beta^\top x_i) - \exp(\beta^\top x_j)|\right) \quad (8)$$

where $C_\triangle(x_i, x_j) = |C_{x_i} - C_{x_j}|$ measures the concordance difference between x_i and x_j. The concordance of individual x_g within the ranking list, C_{x_g}, is defined as:

$$\begin{aligned} C_{x_g} &= \frac{1}{\sum_{g' \neq g} \mathbb{1}[\delta_< = 1]} \sum_{g' \neq g} \mathbb{1}[h(t|x_>) < h(t|x_<), \delta_< = 1] \\ &= \frac{1}{\sum_{g' \neq g} \mathbb{1}[\delta_< = 1]} \\ &\quad \times \sum_{g' \neq g} \mathbb{1}[\exp(\beta^\top x_>) < \exp(\beta^\top x_<), \delta_< = 1] \end{aligned} \quad (9)$$

where $x_>$ and $x_<$ are the individuals with a longer, *i.e.* $T_> = \max(T_g, T'_g)$, and a shorter, *i.e.* $T_< = \min(T_g, T'_g)$, survival time, and $\delta_<$ is the event indicator of shorter survival time. C_{x_g} can be interpreted as the fraction of all other individuals whose predicted survival times are correctly ordered with x_g considering their actual survival times. The concordance difference effectively adjusts the similarity values defined in Equation (7) by penalizing the cases where one individual's predicted survival time aligns well with that of others, while another individual's does not. 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.

Armed with the similarity matrix Sim_D and $\mathrm{Sim}_{D'}$, we propose the *Fair Normalized Discounted Cumulative Gain (FNDCG@k)*, motivated by learning to rank [6], as a metric for the evaluation of

individual fairness with censorship defined as follows:

$$\text{FNDCG@k} = \frac{1}{N}\sum_{n=1}^{N}\frac{\text{DCG}_{\text{Sim}_D(d_n)}}{\text{DCG}_{\text{Sim}_{D'}(d_n)}} \tag{10}$$

where N is the number of individuals and $\text{DCG}_{\text{Sim}(d_n)}$ is the discounted cumulative gain of d_n defined as:

$$\text{DCG}_{\text{Sim}(d_n)} = \sum_{\text{pos}=1}^{k}\frac{\text{Sim}_{D'}(d_{l_{\text{pos}}}, d_n)}{\log(\text{pos}+1)} \tag{11}$$

where k is the length of the ordering list, $\{l_{\text{pos}}\}_{\text{pos}=1}^{k}$ is the ordering list of individual indices derived from the similarity matrix Sim for individual d_n, and $\text{Sim}_{D'}(d_{l_{\text{pos}}}, d_n)$ is the similarity in the input space between the individual at the pos-th position of the ordering list, $d_{l_{\text{pos}}}$, and the individual d_n. It is important to note that both $\text{DCG}_{\text{Sim}_D(d_n)}$ and $\text{DCG}_{\text{Sim}'_D(d_n)}$ are computing the DCG of the similarity values from $\text{Sim}_{D'}$, and the corresponding similarity is used only for deriving the ordering list l_{pos}.

Essentially, FNDCG@k computes the per-individual average ratio between the DCG of input space similarity ranked by output space similarity and the maximum achievable DCG of input space similarity. Therefore, the value of FNDCG@k is within the interval of [0,1], which aligns with the existing individual fairness notions. In addition, the higher the FNDCG@k score, the more consistency between the ranking list encoded from the input and output space and thus, the fairer the model.

The intuition behind enforcing FNDCG@k lies in promoting the consistency between the two ranking lists from the input and output spaces, i.e. having individuals ranked closer in the input space (*e.g.* similar clinical condition) ranked closer in the output space (*e.g.* similar risks and thus similar allocation of medical resources). Moreover, focusing on the top-k ranking promotes local similarity without enforcing global similarity, corresponding to the individual fairness concept of promoting similar outcomes for similar individuals instead of for a group of individuals. Finally, it is worth mentioning that our ranking perspective of individual fairness also possesses the potential of generalization to applications with full label availability, leaving the possibility of future expansion.

4.2 Individual Fairness Algorithm under Uncertainty

With the tailored individual fairness definition accounting for censoring, we now introduce a corresponding learning algorithm, *fairIndvCox*, following the classic Cox proportional hazard model for modeling censored data, to generate tailored forecasts while providing fair risk predictions across individuals. Essentially, the learning algorithm augments the partial likelihood maximization of the CPH model with our individual fairness quantification, FNDCG@k.

Starting with the model utility maximization, the utility loss function $\mathcal{L}_{\text{utility}}$ is formulated as the negative log partial likelihood of the CPH model. Given the partial likelihood in Equation (5), we have defined $\mathcal{L}_{\text{utility}}$ as

$$\mathcal{L}_{\text{utility}} = -\sum_{i:\delta_i=1}(\beta^\top x_i - \log\sum_{j:T_j\geq T_i}\exp(\beta^\top x_j)) \tag{12}$$

Next, we integrate Equation (10) as the individual fairness regularizer $\mathcal{L}_{\text{fairness}} = \text{FNDCG@k}$ and define the unified objective function as

$$\mathcal{L}_{\text{unified}} = \mathcal{L}_{\text{utility}} - \gamma\mathcal{L}_{\text{fairness}} \tag{13}$$

where γ is the tuning parameter controlling the trade-off between utility and fairness. Combining $\mathcal{L}_{\text{utility}}$ and $\mathcal{L}_{\text{fairness}}$ in the unified objective function enables the learned model to be both accuracy-driven and fairness-oriented.

There are two hyper-parameters governing fairIndvCox: γ, the coefficient controlling the balance between utility and fairness, and k, the length of the ordered list in the computation of $\text{DCG}_{\text{Sim}(d_n)}$. Both parameters effect our algorithm as a trade-off between the predictive performance and individual fairness, as we show empirically in Section 5.5 and 5.6.

5 Experiments

We conduct experiments to evaluate the effectiveness of our fairIndvCox algorithm, conduct a comparison study on our Lipschitz-free bias quantification, and examine the trade-offs controlled by the algorithm's hyper-parameters.

5.1 Datasets

We validate our model on four real-world censored datasets with socially sensitive concerns: i) The *ROSSI* dataset pertains to persons convicted then released from Maryland state prisons, who were followed up for one year after release [17]. ii) The landmark algorithmic unfairness *COMPAS* dataset to predict recidivism from Broward County [1]. iii) The *KKBox* dataset from the WSDM-KKBox's Churn Prediction Challenge 2017 [26]. iv) The *Support* dataset of hospitalized patients from five tertiary care academic centers [25]. See Table 1 for the statistics. Note that survival information is explicitly included in these datasets to account for censoring.

Table 1: Summary of datasets used in experiments

Characteristics \ Dataset	ROSSI	COMPAS	KKBox	Support
Sample #	432	10,325	2,814,735	8,873
Censored #	318	7,558	975,834	2,840
Censored Rate	0.736	0.732	0.347	0.320
Feature #	9	14	18	14

5.2 Experiment Setup

Baselines We compare fairIndvCox against four baselines: i) the recently proposed fair survival model FDCPH [24], which, to the best of our knowledge, is the only work for fair survival analysis problem across individuals, ii) the classic survival analysis model CPH [12], iii) the state-of-the-art random forests modeling censored data RSF [20], and iv) the deep neural network on survival analysis DeepSurv [23]. Other competing fairness methods are not considered as none of them is capable of addressing fairness in the presence of censoring. Neither are group-based fair survival models as they necessitate the specification of sensitive attribute to enforce fairness, which is absent in individual fairness learning.

Predictive Performance Measures In addition to the proposed individual fairness measure, we also follow [44] to report survival analysis metrics including C-index, Brier score, and time-dependent AUC as measures of predictive performance under censorship. The C-index [19] evaluates the probability that the predicted event-times of two individuals have the same relative order as their true event-times. The Brier score [5] measures the mean squared difference

Dataset	Method \ Metrics	FNDCG@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
	fairIndvCox	**53.29** (**20.78%**)	63.78 (-4.34%)	15.12 (-2.79%)	78.25 (-1.35%)
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
	fairIndvCox	**83.14** (**11.39%**)	68.73 (-8.50%)	13.97 (-4.10%)	**71.87** (0.05%)
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
	fairIndvCox	**67.44** (**15.01%**)	**83.27** (0.31%)	14.45 (-1.47%)	**80.95** (0.29%)
Support	FDCPH	62.31	67.88	30.54	76.34
	CPH	55.78	74.11	21.21	80.02
	RSF	65.15	75.18	16.64	**81.01**
	DeepSurv	54.33	**75.65**	**16.11**	80.68
	fairIndvCox	**72.17** (**10.78%**)	74.31 (-1.16%)	17.13 (-6.33%)	79.51 (-1.85%)

Table 2: Evaluation results of different models with the best results marked in bold. The numbers in parentheses represent the relative performance improvement of fairIndvCox compared to the best baseline.

between the predicted probability of outcome assignments and the true outcomes (the lower the Brier score, the better the prediction). The time-dependent AUC [8] quantifies the probability that a randomly selected pair of individuals having experienced the event and not having experienced the event at time t are correctly ordered.

Without loss of generality, we employ the Euclidean distance with feature scaling to obtain $\text{Sim}_{D'}$. All methods are trained in the same way with 5-fold cross validation for fair comparison. The Adam optimizer is used to optimize the model via backpropagation and automatic differentiation in PyTorch, with a learning rate of 0.01. The training is done in mini-batches of size 128 for 50 epochs. The overall objective function for quantitative performance comparison has top k set as 10 and γ set as 1. The base model Cox's hyperparameter settings are followed for the hidden unit number, and a grid search is conducted for fairness-specific tuning parameters. The search space for k is 4-50 and for γ it is e^{-4} and e^{4}.

5.3 Experiment Results

Table 2 shows the results of our experiment. Our new fairIndvCox clearly dominates all other baselines in minimizing discrimination (measured by FNDCG@10) while maintaining a competitive predictive performance (measured by C-index, Brier score, and time-dependent AUC), which verifies the necessity of its debiasing design across individuals while accounting for censorship. In contrast, the compared methods suffer from the lack of considering censored data as well as the non-trivial handling of Lipschitz constant. The improved overall predictive performance of fairIndvCox also shows the merit of such an anti-discrimination design for prediction accuracy, presumably due to fairness regularization reducing overfitting.

5.4 Comparison Study on the Lipschitz-free Bias Quantification

We further perform a comparison study to demonstrate our method's advantage brought by being free from the Lipschitz condition in Equation (1) which requires the specification of the Lipschitz constant during fairness quantification. We replace the $\mathcal{L}_{\text{fairness}}$ in fairIndvCox with Equation (1) as suggested in [15], and denote the method as *fairIndvCox-*. Results in Table 3 show that fairIndvCox outperforms fairIndvCox- in minimizing discrimination for all datasets by large margins and also in terms of the predictive performance, except for a small decrease in the ROSSI dataset. This verifies that relaxing the Lipschitz constant specification in the conventional individual fairness definition can lead to improved performance.

Table 3: Results of comparison study on the Lipschitz-free Bias Quantification.

Dataset	FNDCG@10%		C-index%	
	fairIndvCox-	fairIndvCox	fairIndvCox-	fairIndvCox
ROSSI	45.29	53.29	64.42	63.78
COMPAS	77.39	83.14	60.14	68.73
KKBox	54.02	67.44	82.71	83.27
Support	58.49	72.17	69.28	74.31

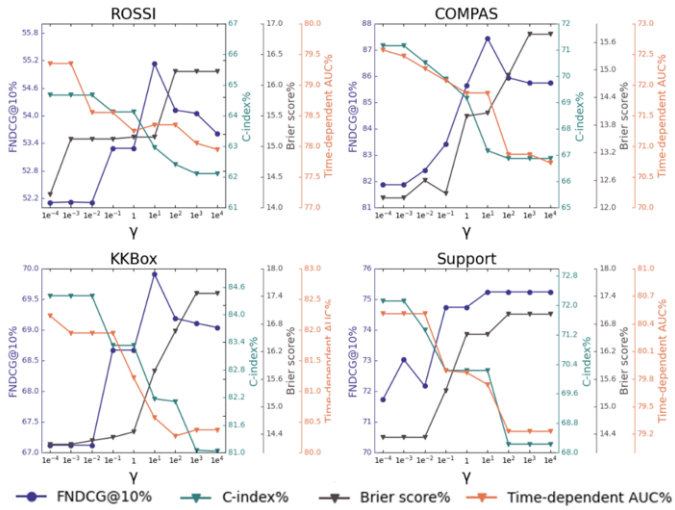

Figure 2: Study on individual fairness and accuracy trade-off on γ: The fairIndvCox models subject to different γ variations (between e^{-4} and e^{4}) on ROSSI, COMPAS, KKBox, and Support exhibit effects on individual fairness and model accuracy.

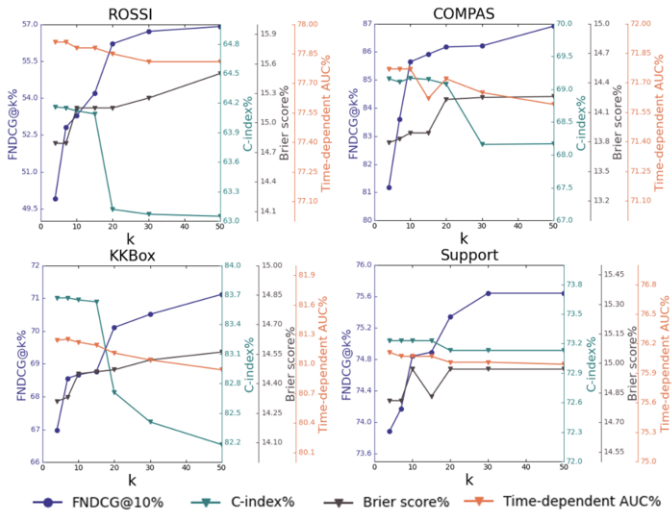

Figure 3: Study the choice of k-value: The fairIndvCox models subject to different k variations (between 4 and 50) on ROSSI, COMPAS, KKBox, and Support exhibit effects on individual fairness and model accuracy.

5.5 *Effect of Different γ Values on Individual Fairness and Predictive Performance*

To investigate the effect of γ on the performance of fairIndvCox, we vary γ within the set $\{e^{-4}, e^{-3}, \cdots, e^{4}\}$ where e is the natural constant, keeping all other hyper-parameters unchanged. We compare fairIndvCox's performance in terms of predictive power and individual fairness under the different settings.

According to the results shown in Figure 2, there are three cases of γ values. (1) For small γ (*i.e.*, less than e^{-2} for ROSSI, e^{-3} for COMPAS, e^{-2} for KKBox, and e^{-3} for Support), the individual fairness constraint has a small effect on fairIndvCox's predictive performance metrics (C-index, Brier score, and time-dependent AUC) and FNDCG@10 for the four tasks. (2) As γ increases progressively (*e.g.*, from e^{-2} to e^{1} for ROSSI, e^{-3} to e^{1} for COMPAS, e^{-2} to e^{1} for KKBox, and e^{-3} to e^{1} for Support), fairness increases significantly but at the cost of some predictive performance degradation (decreased C-index, decreased time-dependent AUC, and increased Brier score). This would imply that fairIndvCox achieves the appropriate balance between fostering individual fairness and preserving model performance. (3) When γ is relatively large (*e.g.*, larger than e^{1} for all the datasets), the promotion of individual fairness will continue to have an effect on the predictive performance, with the exception of the Support dataset where both FNDCG@10 and the predictive performance metrics stay mostly fixed when γ is greater than e^{2}. Note that FNDCG@10 mostly also decreases as γ increases since it adds more weight to $\mathcal{L}_{\text{fairness}}$. But this does not mean we can obtain the optimal node. Therefore, the performance of individual fairness promotion within a fixed epoch is close to its limit, and it is difficult to achieve better performance.

5.6 *Effect of Different Number of Neighbors k Values on Individual Fairness and Predictive Performance*

Similar to the previous section, we conducted experiments with a variety of values for k in $\{4, 7, 10, 15, 20, 30, 50\}$, keeping all other training factors the same. We compare fairIndvCox's predictive performance and fairness under different settings.

We observe that (Figure 3): (1) As k increases, the fairIndvCox achieves better performance on *FNDCG@k*, demonstrating better optimization for individual fairness. (2) When k is a modest value (e.g., smaller than 15 for ROSSI, 20 for COMPAS, 15 for KKBox, and 10 for Support), the predictive performance (as measured by C-index and time-dependent AUC) is hardly affected or even increases, though the Brier score performs slightly worse (increased). The fairIndvCox mostly strikes the right balance between maintaining model utility and fostering individual fairness with proper choices of k in here. (3) When k is significant (e.g., greater than 15 for ROSSI, 20 for COMPAS, 15 for KKBox, and 10 for Support), the predictive performance significantly declines (decreased C-index and time-dependent AUC, and increased Brier score), with the exception on the Support dataset where the Brier score fluctuates when k is between 10 and 20, and all three metrics stay relatively flat when k is greater than 20. In general, more points will be referenced at a time as k increases, resulting in more interference values. This leads to a decrease in the weight of the correct label and a blurred classification, causing degradation in predictive performance.

6 Conclusion

A striking gap exists between the prevailing real-world applications with censorship and the assumption of full class label availability in existing AI fairness methods. We make an initial investigation of individual fairness in learning with censorship. Moreover, this work defines individual fairness from a ranking perspective, relaxing from the Lipschitz condition in conventional individual fairness studies. The proposed notion and algorithm are expected to be versatile in quantifying and mitigating bias in various socially sensitive applications. We provide an empirical evaluation of four real-world datasets to validate the effectiveness of our method. The experimental results show that with suitable γ and k values, our method can substantially improve individual fairness with an acceptable loss of predictive performance as the model outperforms the current state-of-the-art individual fairness promotion methods. Finally, this work defines a new task that opens up possibilities for future work to achieve more applicable and comprehensive AI fairness.

References

[1] J Angwin, J Larson, S Mattu, and L Kirchner, 'There's software used across the country to predict future criminals', *And it's biased against blacks. ProPublica*, (2016).

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan, 'Fairness in machine learning', *Nips tutorial*, **1**, 2, (2017).

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, et al., 'Putting fairness principles into practice: Challenges, metrics, and improvements', *AIES*, (2019).

[4] Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, et al., 'A review of survival trees', *Statistics surveys*, **5**, 44–71, (2011).

[5] Glenn W Brier and Roger A Allen, 'Verification of weather forecasts', in *Compendium of meteorology*, 841–848, Springer, (1951).

[6] Christopher Burges, Robert Ragno, and Quoc Le, 'Learning to rank with nonsmooth cost functions', *NIPS*, **19**, (2006).

[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy, 'Building classifiers with independency constraints', in *ICDMW*, pp. 13–18, (2009).

[8] Lloyd E Chambless and Guoqing Diao, 'Estimation of time-dependent area under the roc curve for long-term risk prediction', *Statistics in medicine*, **25**(20), 3474–3486, (2006).

[9] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi, 'Ethical machine learning in healthcare', *Annual Review of Biomedical Data Science*, **4**, (2020).

[10] A Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.', *Big Data*, **5**(2), 153–163, (2017).

[11] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman, 'Survival analysis part i: basic concepts and first analyses', *British journal of cancer*, **89**(2), 232–238, (2003).

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

[13] Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li, 'Individual fairness for graph neural networks: A ranking based approach', in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 300–310, (2021).

[14] Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu, 'Learning credible dnns via incorporating prior knowledge and model local explanation', *Knowledge and Information Systems*, **63**(2), 305–332, (2021).

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, et al., 'Fairness through awareness', in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, (2012).

[16] Benjamin Fish, Jeremy Kun, et al., 'A confidence-based approach for balancing fairness and accuracy', in *SDM*, pp. 144–152, (2016).

[17] John Fox, Marilia S Carvalho, et al., 'The rcmdrplugin. survival package: Extending the r commander interface to survival analysis', *Journal of Statistical Software*, **49**(7), 1–32, (2012).

[18] Moritz Hardt, Eric Price, Nati Srebro, et al., 'Equality of opportunity in supervised learning', in *Advances in neural information processing systems*, pp. 3315–3323, (2016).

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

[20] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al., 'Random survival forests', *Annals of Applied Statistics*, **2**(3), 841–860, (2008).

[21] Faisal Kamiran and Toon Calders, 'Classifying without discriminating', in *2nd International Conference on Computer, Control and Communication*, pp. 1–6, (2009).

[22] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong, 'Inform: Individual fairness on graph mining', in *Proceedings of the 26th KDD Conference*, pp. 379–389, (2020).

[23] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger, 'Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network', *BMC medical research methodology*, **18**(1), 1–12, (2018).

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

[25] William A Knaus, Frank E Harrell, Joanne Lynn, et al., 'The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults', *Annals of internal medicine*, **122**(3), 191–203, (1995).

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

[27] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum, 'ifair: Learning individually fair data representations for algorithmic decision making', in *2019 ieee 35th international conference on data engineering (icde)*, pp. 1334–1345. IEEE, (2019).

[28] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum, 'Operationalizing individual fairness with pairwise fair representations', *Proceedings of the VLDB Endowment*, **13**(4), (2019).

[29] Wenhua Liang, Hengrui Liang, Limin Ou, Binfeng Chen, Ailan Chen, Caichen Li, et al., 'Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19', *JAMA internal medicine*, **180**(8), 1081–1089, (2020).

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*, **54**(6), 1–35, (2021).

[31] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun, 'Two simple ways to learn individual fairness metrics from data', in *International Conference on Machine Learning*, pp. 7097–7107. PMLR, (2020).

[32] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar, 'On ranking in survival analysis: Bounds on the concordance index', *NIPS*, **20**, (2007).

[33] Anthony Joe Turkson, Francis Ayiah-Mensah, and Vivian Nimoh, 'Handling censoring and censored data in survival analysis: a standalone systematic literature review', *International journal of mathematics and mathematical sciences*, **2021**, 1–16, (2021).

[34] Sriram Vasudevan and Krishnaram Kenthapadi, 'Lift: A scalable framework for measuring fairness in ml applications', in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2773–2780, (2020).

[35] Changlin Wan, Wennan Chang, Tong Zhao, Sha Cao, and Chi Zhang, 'Denoising individual bias for fairer binary submatrix detection', in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 2245–2248, (2020).

[36] Ping Wang, Yan Li, and Chandan K Reddy, 'Machine learning for survival analysis: A survey', *ACM Computing Surveys (CSUR)*, **51**(6), 1–36, (2019).

[37] Xuejian Wang, Wenbin Zhang, Aishwarya Jadhav, and Jeremy Weiss, 'Harmonic-mean cox models: A ruler for equal attention to risk', in *Survival Prediction-Algorithms, Challenges and Applications*, pp. 171–183. PMLR, (2021).

[38] Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, and Wenbin Zhang, 'Preventing discriminatory decision-making in evolving data streams', in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 149–159, (2023).

[39] Zichong Wang, Charles Wallace, Albert Bifet, Xin Yao, and Wenbin Zhang, 'FG^{2}AN: Fairness-aware graph generative adversarial networks', in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Turin, Italy, (2023).

[40] Zichong Wang, Yang Zhou, Meikang Qiu, Israat Haque, Laura Brown, Yi He, Jianwu Wang, David Lo, and Wenbin Zhang, 'Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking', *arXiv preprint arXiv:2302.08018*, (2023).

[41] Jiaming Zeng, Berk Ustun, and Cynthia Rudin, 'Interpretable classification models for recidivism prediction', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **180**(3), 689–722, (2017).

[42] Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C Weiss, and Wolfgang Nejdl, 'Farf: A fair and adaptive random forests classifier', in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 245–256. Springer, (2021).

[43] Wenbin Zhang and Eirini Ntoutsi, 'Faht: an adaptive fairness-aware decision tree classifier', in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1480–1486, (2019).

[44] Wenbin Zhang and Jeremy Weiss, 'Longitudinal fairness with censorship', in *Proceedings of the AAAI Conference*, (2022).

[45] Chen Zhao and Feng Chen, 'Rank-based multi-task learning for fair regression', in *2019 IEEE ICDM*, pp. 916–925. IEEE, (2019).

[46] Indre Žliobaite, Faisal Kamiran, and Toon Calders, 'Handling conditional discrimination', in *2011 IEEE 11th International Conference on Data Mining*, pp. 992–1001. IEEE, (2011).